Received: 3 July 2023 | Revised: 10 November 2023 | Accepted: 13 December 2023

DOI: 10.1111/1755-0998.13920

MOLECULAR ECOLOGY RESOURCES WILEY

RESOURCE ARTICLE

A CRISPR-based strategy for targeted sequencing in biodiversity science

Bethan Littleford-Colquhoun[1,2] | Tyler R. Kartzinel[1,2]

[1]Department of Ecology, Evolution, and Organismal Biology, Brown University, Providence, Rhode Island, USA

[2]Institute at Brown for Environment and Society, Brown University, Providence, Rhode Island, USA

Correspondence
Bethan Littleford-Colquhoun and Tyler R. Kartzinel, Department of Ecology, Evolution, and Organismal Biology, Brown University, Providence, RI 02912, USA.
Email: bethan_littleford-colquhoun@brown.edu and tyler_kartzinel@brown.edu

Funding information
Institute at Brown for Environment and Society seed award; National Science Foundation, Grant/Award Number: NSF DEB-2046797

Handling Editor: Carla Martins Lopes

Abstract

Many applications in molecular ecology require the ability to match specific DNA sequences from single- or mixed-species samples with a diagnostic reference library. Widely used methods for DNA barcoding and metabarcoding employ PCR and amplicon sequencing to identify taxa based on target sequences, but the target-specific enrichment capabilities of CRISPR-Cas systems may offer advantages in some applications. We identified 54,837 CRISPR-Cas guide RNAs that may be useful for enriching chloroplast DNA across phylogenetically diverse plant species. We tested a subset of 17 guide RNAs in vitro to enrich plant DNA strands ranging in size from diagnostic DNA barcodes of 1,428 bp to entire chloroplast genomes of 121,284 bp. We used an Oxford Nanopore sequencer to evaluate sequencing success based on both single- and mixed-species samples, which yielded mean chloroplast sequence lengths of 2,530–11,367 bp, depending on the experiment. In comparison to mixed-species experiments, single-species experiments yielded more on-target sequence reads and greater mean pairwise identity between contigs and the plant species' reference genomes. But nevertheless, these mixed-species experiments yielded sufficient data to provide ≥48-fold increase in sequence length and better estimates of relative abundance for a commercially prepared mixture of plant species compared to DNA metabarcoding based on the chloroplast *trn*L-P6 marker. Prior work developed CRISPR-based enrichment protocols for long-read sequencing and our experiments pioneered its use for plant DNA barcoding and chloroplast assemblies that may have advantages over workflows that require PCR and short-read sequencing. Future work would benefit from continuing to develop in vitro and in silico methods for CRISPR-based analyses of mixed-species samples, especially when the appropriate reference genomes for contig assembly cannot be known a priori.

KEYWORDS
amplicon sequencing, barcoding, Cas9, contigs, environmental DNA, Flye, guide RNA, long-read sequencing, metabarcoding, metagenomics

1 | INTRODUCTION

In biomedical science, CRISPR-Cas systems are regularly used to target a section of DNA with high precision and accuracy (Kaminski et al., 2021; Wang et al., 2022). Although most applications of CRISPR have utilized its genome-editing capabilities, its target-specific binding and cutting capabilities for DNA detection and enrichment are increasingly evident (Phelps et al., 2020). To date, CRISPR has been used: to detect the presence of specific genes of interest such as antibiotic resistance in *Staphylococcus* (Quan et al., 2019), drug resistance in the

malaria parasite *Plasmodium falciparum* (Cunningham et al., 2021), cancer cell lines (Stangl et al., 2020), and SARS-Cov-2 (Broughton et al., 2020); to identifying SNPs associated with lung cancer (Qiu et al., 2018), the hepatitis B virus (Ke et al., 2021), and bacterial genes within environmental samples (Sandoval-Quintana et al., 2023). While CRISPR-Cas systems are becoming the most reliable, affordable and versatile method for analysing nucleic acids, they may be generally underutilized in environmental biology (Phelps et al., 2020).

The main components of CRISPR-Cas systems that may be useful for applications requiring sequence-based taxonomic identifications are conceptually very similar to those that are widely used today in PCR-based methods for DNA barcoding and metabarcoding. Type II CRISPR-Cas systems are the best characterized and most commonly used (Xu & Li, 2020) and comprise of two key components: a guide RNA (gRNA), which recognizes the target sequence with high precision (Knott & Doudna, 2018) and a CRISPR-associated endonuclease (Cas protein) that cuts the targeted sequence. Guide RNAs are composed of a 'scaffold sequence' necessary for Cas-binding and a user-defined ~20 nucleotide 'spacer sequence' that correspond to a 'target sequence' to be cleaved from template DNA by the Cas-gRNA ribonucleoprotein (RNP) complex (López-Girona et al., 2020). Much like designing PCR primers, specific sequences can be targeted based on the user-defined spacer sequence of the gRNA. Similarly, the gRNA can tolerate some degree of mismatching between target and spacer sequences: the gRNA binds to the target in a 3′ to 5′ direction such that mismatches at the 3′ end of the spacer can prevent cleavage whereas ~2 bp mismatches towards the 5′ end may often be tolerated (Fu et al., 2016). Conveniently, CRISPR gRNAs can also be ordered from the same manufacturers as the oligos used as PCR primers. What makes CRISPR so reliable and versatile is its ability to recognize a target sequence with high precision (Knott & Doudna, 2018).

Despite the many similarities between PCR and CRISPR-Cas systems for detecting and identifying DNA sequences, there are some important differences. Although gRNAs are conceptually similar to PCR primers when used to enrich a sample for target sequences, the process does not amplify copies of the target as in PCR but rather cleaves the target from a genomic sample in proportion to its abundance. To design CRISPR-Cas assays for enrichment, a target sequence must be located immediately before a nuclease-specific 'protospacer adjacent motif' (PAM) and fortunately these are numerous throughout the genome. A distinct benefit to using CRISPR-Cas enrichment is that, unlike PCR, many (≥100) gRNAs can be multiplexed within a single assay (Gilpatrick et al., 2023; López-Girona et al., 2020; Xie et al., 2015), allowing for multiple regions within a genome to be enriched and potentially assembled in a single reaction. Multiple scoring models have been developed to help identify efficient and specific gRNAs which integrate the assessment of GC count and thermodynamic properties (both of which are often used to assess PCR primer design), as well as position-independent nucleotide counts and the location of the gRNA target site within the gene. The commonly used scoring methods, however, vary in their intended uses and thus it has been historically challenging to translate their utility beyond model systems (Cui et al., 2018; Sledzinski et al., 2020; Wilson et al., 2018). As the ability to design effective gRNAs continues to improve, many more potential applications in environmental biology may begin to be realized (Gilpatrick et al., 2023).

Clearly, methods developed using PCR or CRISPR-Cas may have complementary strengths and weaknesses with applications in environmental biology. Standard DNA barcoding and metabarcoding methods rely on PCR to enrich sequences from single- or mixed-species samples, respectively, in order to compare the resulting sequences with reference data (Srivathsan et al., 2021). Unfortunately, existing DNA reference databases are often biased towards certain markers and it has been difficult to achieve consensus about which barcodes to use for certain taxa, in part because reliance on PCR limits the length of target sequences in ways that can constrain taxonomic precision (CBOL Plant Working Group, 2009; Hebert et al., 2022; Hoban et al., 2022; Keck et al., 2022). When DNA metabarcoding approaches are applied to samples containing mixtures of DNA from multiple species, reliance on PCR involves further challenges associated with detecting and estimating the relative abundance of phylogenetically disparate taxa (Clarke et al., 2014; Deagle et al., 2019; Kelly et al., 2019; O'Donnell et al., 2016; Stapleton et al., 2022). By contrast, CRISPR-Cas systems may enable researchers to circumvent several of these challenges and overcome drawbacks to PCR by providing longer and hence more diagnostic markers. Recent CRISPR applications in environmental biology have already exemplified its versatility by detecting specific DNA strands in environmental DNA (Baerwald et al., 2023; Karlikow et al., 2023; Sánchez et al., 2022; Shashank et al., 2023; Williams et al., 2019, 2021, 2023), enabling the targeted enrichment of fish mitogenomes (Ramón-Laca et al., 2023), and identifying structural variants in loci controlling for colour in apples (López-Girona et al., 2020). However, it has not yet been used to study the chloroplast genomes of plants or used in comparative studies involving multiple target sequences within a single sample.

We developed a set of novel CRISPR-based protocols to enrich plant DNA barcodes. We began by evaluating the availability of gRNA sequences capable of targeting chloroplast DNA across a broad swath of the angiosperm phylogeny, which should enable 'universal' DNA enrichment strategies. Then we designed protocols to target the enrichment of CRISPR-associated loci in vitro. We evaluated the strengths and weaknesses of broad-spectrum strategies for enriching markers that ranged in size from 1,428 bp to the entire chloroplast genome from single- or mixed-species samples by: (*i*) comparing three strategies for enriching standard plant DNA barcode loci from a single-species DNA sample, (*ii*) enriching a whole chloroplast to assemble a reference genome for a single species and (*iii*) applying a barcode-enrichment strategy to a mixed sample of known species composition. The strengths and weaknesses we report from each experimental approach will help inform future assay development and further research as required to better understand the challenges and opportunities that each type of experiment may present in the field.

2 | METHODS

2.1 | Assessing cross-species coverage of guide RNAs (gRNAs)

Our goal was to identify broad-spectrum gRNAs that targeted chloroplast DNA sequences from many species. We designed gRNAs (20 bp in length) for the Type II CRISPR-Cas system. This system relies on the Cas9 (SpCas9) protein which recognizes a PAM sequence of NGG in a 5′ to 3′ direction (where 'N' can be any nucleotide base). We began by identifying candidate gRNAs that appeared in chloroplast reference genomes across a set of 7 well-studied, economically important and phylogenetically disparate flowering plant species: three grasses (wheat, *Triticum aestivum*; oats, *Avena sativa*; corn, *Zea mays*), two superrosids (soybeans, *Glycine max*; peanuts, *Arachis hypogaea*) and two superasterids (sunflower, *Helianthus annuus*; spinach, *Spinacia oleracea*; see Table 2 for RefSeq accession numbers). We did this by searching for all potential gRNAs in the chloroplast reference genomes of the 7 target species using the *Find CRISPR site* tool within Geneious Prime 2023.0.4. We evaluated the predicted in vitro functionality of these gRNAs based on features including GC count, position-independent nucleotide counts, the location of the gRNA target site within the gene and the thermodynamic properties of each identified gRNA using the *Rule Set 2* scoring method (Doench et al., 2016). The *Rule Set 2* model gives high scores to candidate gRNAs that are predicted to efficiently guide Cas9 to the correct spot for cleavage (i.e., on-target activity), enabling comparisons of the candidate gRNAs across genomic sites and target taxa. Once candidate gRNAs were identified using each reference genome independently, we tallied the number of references that contained an exact (100%) match between the guide and the target sequences (i.e., assuming strict, no tolerance for mismatches). We evaluated (*i*) how many exact gRNAs were present in only one species (i.e., narrowest coverage), (*ii*) how many unique gRNAs occurred across all species (i.e., broadest coverage), and (*iii*) how many gRNAs had multiple match sites within a species (i.e., poor site fidelity). Then we identified gRNAs that exactly matched ≥5 of the 7 target species, tolerating up to 2 mismatches at the 5′ end of the gRNA. Finally, we selected candidate gRNAs that had good predicted in vitro functionality and coverage across the 7 target species based on the criteria outlined above and had a *Rule Set 2* score ≥0.2 in all target species.

2.2 | Selection of gRNAs and in vitro testing

We selected a subset of broad-coverage candidate gRNAs for use in a series of six experiments to: (*i*) sequence standard plant DNA barcode loci and (*ii*) a complete chloroplast genome from a DNA sample representing a single species (spinach) as well as to (*iii*) elucidate the sequence composition of a mixed-sample containing six known plant species (wheat, oats, corn, soybean, peanuts, sunflower; Figure 1). We refer to these overarching strategies, intuitively, as the 'barcoding approach', 'whole chloroplast approach' and 'mixed-species approach'.

We began with the relatively simple, single species 'barcoding approach' using spinach as the target species (Experiments 1–3; Figure 1). We targeted three 'standard' plant DNA barcodes as well as other markers that have been considered potentially useful for DNA barcoding (CBOL Plant Working Group, 2009; Kress, 2017). Experiment 1 used 2 gRNAs to target the whole *rbc*L gene (1,428 bp), which includes the standard *rbc*L barcode locus (553 bp; CBOL Plant Working Group, 2009; Kress, 2017). One gRNA targeted a region upstream of the barcode (the 'forward' gRNA) and one targeted a region downstream of the barcode ('reverse') such that the two gRNA binding sites were separated by 5,809 bp (Figure 1, Table 1). Experiment 2 targeted multiple DNA barcodes by making a two-directional break in the *trn*G gene with forward and reverse gRNAs that overlapped by 18 bp (Figure 1, Table 1). Within spinach, four potentially useful plant DNA barcodes sit within 9,000 bp of *trn*G and can be targeted for sequencing in this way (the standard *mat*K and *trn*H-*psb*A barcodes as well as *psb*K-*psb*I and *atp*F-*atp*H; CBOL Plant Working Group, 2009; Kress, 2017). Experiment 3 targeted the inverted 16S rRNA repeat region (1,491 bp) of the chloroplast genome and aimed to determine whether we could use a single gRNA to sequence both regions, because 16S is a structurally interesting region of the chloroplast (Manhart, 1995; Strauss et al., 1988) even though it is more often targeted as a DNA barcode for other taxa (e.g., bacteria [Caporaso et al., 2012], animals [Kartzinel & Pringle, 2015; Vences et al., 2005]; Figure 1, Table 1).

Our second overarching aim was to test methods suitable for a 'whole chloroplast approach' using spinach as the target species (Experiment 4, Figure 1). We used 12 gRNAs that were predicted to enable enrichment around 20 cut sites that were relatively evenly spaced throughout the spinach chloroplast (~5,200–17,500 bp apart). Of the 12 gRNAs that we selected, 8 occurred only once in the spinach reference chloroplast (forward in directionality; t1, t2, t4, t6, t7, t8, t9, t13), 2 occurred twice (only when mismatches were tolerated; t10, t16) and 2 occurred 4 times as they were located in the large inverted rRNA subunit repeats (t12, t14; Figure 1, Table 1).

Finally, we selected candidate gRNAs to test a 'mixed-species approach' for identifying taxa (Experiments 5–6; Figure 1). These experiments aimed to sequence 6 target species that were ground and pelleted by the commercial supplier Teklad Lab Animal Diets. These pellets represented a homogeneous mixture comprising Teklad Global Rodent 2016 formula (wheat, corn, soybean; TD.00217) and 3 additional plant components (oats, peanuts, sunflowers) that were mixed in even biomass proportions. The overall biomass ratios were 50% Teklad Global Rodent 2016 to 50% additional plant components, resulting in hypothetical plant biomass ratios of oat, peanut and sunflower at 1/6th each plus wheat, corn, and soy that each comprised an unspecified ratio within the other 50%; because the core Teklad components were

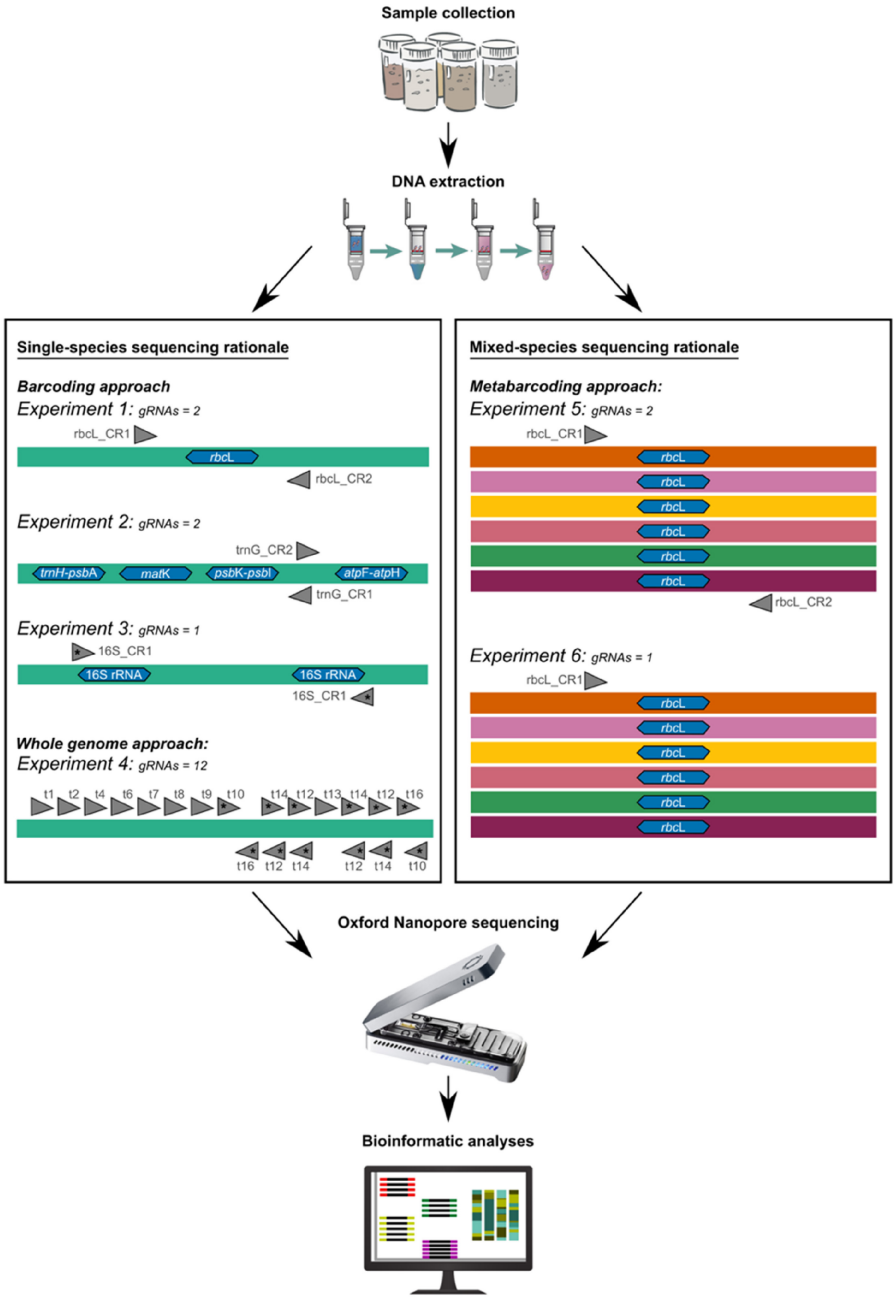

FIGURE 1 Experimental overview for single- and mixed-species sequencing approaches. All experiments began with sample collection and DNA extraction (top) and ended with Oxford Nanopore MinION sequencing followed by bioinformatic analyses (bottom). Experiments differed according to the strategy for designing gRNAs used for enrichment (box; a number of unique gRNAs used are shown for each experiment). The coloured bands represent DNA strands from each of the seven species used across these experiments, the blue hexagons identify the barcode markers targeted in each experiment and grey arrowheads show the location and directionality of each gRNA binding site (Table 1). Asterisks at the biding sites indicate gRNAs that occur at ≥2 locations within the chloroplast, as expected for the inverted repeats of 16S rRNA.

processed more intensively than the additional components we assumed a greater level of DNA degradation in the former than the latter. Experiments 5 and 6 both targeted *rbc*L, but Experiment 5 used both a forward and reverse gRNA while Experiment 6 used only a forward gRNA (Figure 1, Table 1). Due to chloroplast rearrangements (Li et al., 2016), the gRNAs appear at different genomic locations across the six reference chloroplast genomes but are linked to the *rbc*L locus in each. We compared Experiments 5 and 6 to determine which provided a greater number of on-target sequences and a better estimate of plant DNA relative read abundance (RRA) as expected based on the biomass of taxa incorporated into the mixture.

TABLE 1 Guide RNA (gRNA) sequences used to direct CRISPR-Cas9 scission in Experiments 1–6.

gRNA	Sequence 5′-3′	Direction	Position 5′-3′	Experiment
*rbc*L_CR1	ACTCTCATACGAGCTCCCGG	Forward	52294–52313	1, 5, 6
*rbc*L_CR2	GGAAAGACTAGGCCTACTAA	Reverse	58142–58123	1, 5
*trn*G_CR1	TCGTTAGCTTGGAAGGCTAG	Reverse	8908–8889	2
*trn*G_CR2	AGCCTTCCAAGCTAACGATG	Forward	8891–8910	2
16S_CR1	ATTAGCTCTCCCTGAAAAGG	Forward; Reverse	133993–134012; 99452–99433	3
sgRNA_t1	TCTCTCTAAAATTGCAGTCA	Forward	1258–1277	4
sgRNA_t2	GCAGTACCTTGACCAACTCC	Forward	12639–12658	4
sgRNA_t4	CAGCTTCCGCCTTGACAGGG	Forward	29129–29148	4
sgRNA_t6	GCCATATTATTAAAAGCTTG	Forward	42076–42095	4
sgRNA_t7	ATTGGTTCAAATCCAATAGT	Forward	50907–50926	4
sgRNA_t8	AGGAATTCTTCCAGTAGTAT	Forward	62520–62539	4
sgRNA_t9	ACTCGTTATCAATGGGATCA	Forward	71857–71868	4
sgRNA_t10	TCTCCAATTATAGCCCCTCT	Forward; Reverse	83418–83437; 150008–150027	4
sgRNA_t12	GCTCTACCACTGAGCTACTG	Forward; Reverse	106007–106026; 127438–127419	4
sgRNA_t13	GGACGAATTTTCCATCTCCA	Forward	119819–119910	4
sgRNA_t14	TAGCTCAGTGGTAGAGCGGT	Forward; Reverse	127422–127441; 106023–106004	4
sgRNA_t16	TGATTGTCTGATAATGAGCA	Forward; Reverse	144744–144763; 88701–88682	4

Note: For each gRNA, we provide a unique identifier, the sequence, the direction of activity ('forward' direction indicates that the protospacer adjacent motif (PAM) and target sequence is found upstream of the region of interest on the forward DNA strand; 'reverse' indicates that the PAM and target sequence is found downstream of the region of interest on the reverse DNA strand), the position of the gRNA with respect to the spinach reference chloroplast genome (Table 2), and the experiment(s) for which we trialled each gRNA (Figure 1).

2.3 | Sequencing library preparation

Total genomic DNA was extracted from 5 replicates of ~0.2 mg (*i*) spinach and (*ii*) mixed-species Teklad samples using a Zymo Quick-DNA Fecal/Soil Microbe Miniprep Kit (Zymo Research). Following extractions, we quantified DNA using a Qubit dsDNA high-sensitivity assay kit (Invitrogen). For each experiment, we enriched the target chloroplast regions using the nanopore Cas9-targeted sequencing method (nCATS; Gilpatrick et al., 2020). Briefly, this method uses Cas9-mediated DNA cleavage to cut double-stranded DNA ~3–4 nucleotides upstream of the target PAM sequence. This enables us to enrich target DNA by selectively ligating adapters to the cut sites created by the Cas9/gRNA–ribonucleoprotein (RNP) complexes created in each experiment. We built custom gRNA duplexes for each experiment and assembled them into the RNP complex by adding 1 μL of pooled crRNAs (user-defined spacer sequences; IDT) and 1 μL of Alt-R® CRISPR-Cas9 tracrRNA (IDT) to 8 μL of nuclease-free water and incubating at 95°C for 5 min. The RNP complex was then created by incubating 1.2 μL Alt-R® HiFi Cas9 Nuclease (IDT), 2.8 μL 10× CutSmart Buffer (NEB), 23 μL nuclease-free water and 3 μL of the gRNA duplex at room temperature for 20 min. To selectively enrich the region of interest, we first dephosphorylated pre-existing DNA ends before cutting with Cas9 to preferentially ligate sequencing adapters to the cut sites created by the RNP complex (Gilpatrick et al., 2020). We did this by incubating 1.5 ng of genomic DNA, 3 μL 10× CutSmart buffer and 3 μL QuickCIP enzyme (NEB) at 37°C for 10 min, followed by enzyme inactivation at 80°C for 2 min. Cleavage and dA-tailing of the dephosphorylated DNA occurred in a reaction using 10 μL of the assembled RNP complex, 10 mM dATP (Zymo Research), and 1 μL Taq DNA polymerase (NEB) with incubation at 37°C for 15 min followed by 72°C for 5 min.

2.4 | Oxford Nanopore sequencing

For each experiment, we sequenced the enriched target loci used long-read nanopore sequencing (Gilpatrick et al., 2020). Nanopore sequencing adapters were first ligated to Cas9 cut sites by incubating 10 μL of NEBNext Quick T4 DNA Ligase (NEB), 20 μL ligation buffer (ONT), 4.5 μL nuclease-free water, and 3.5 μL AMX sequencing adapters (LSK109 sequencing kit; ONT) at room temperature for 10 min. An equivolume amount of TE buffer was then added to the ligated sample, followed by a 0.3× volume addition of AMPure XP beads (Beckman Coulter). The sample was incubated at room temperature for 5 min. The supernatant was removed by pipette after placing the sample on a magnetic rack and then the remaining library was purified twice using 200 μL long-fragment buffer (ONT). We eluted the ligated sample by adding 15 μL elution buffer and incubating at room temperature for 30 min before separating the eluate from beads on the magnetic rack. To ensure the recommended 5–50 fmol of library DNA was available for sequencing, we checked library concentrations with a Qubit dsDNA high-sensitivity kit. The resulting libraries were sequenced on a MinION Mk1B Nanopore sequencer (ONT) using FLO-MIN106D

(R9.4.1) flow cells. We added 37.5 μL sequencing buffer (ONT) and 25.5 μL loading beads to the eluate and then prepared the flow cell by placing 30 μL flush tether (ONT) into a tube of flush buffer (ONT), pulling 230 μL buffer from the priming port, and loading an initial 800 μL of the priming mix. After 5 min, an additional 200 μL of priming mix was loaded before the DNA library. The DNA library was added via the SpotON sample port in a dropwise fashion. Finally, we initiated sequencing runs using MinKNOW software (version 22.08.9; ONT), enabling raw data to be processed with *fast basecalling* using Guppy 6.2.11.

2.5 | Oxford Nanopore read assembly

First, adapters were trimmed from all reads that passed the Guppy basecaller quality score ($Q \geq 8$) using Porechop (Wick et al., 2017). To obtain consensus sequences from overlapping reads, trimmed reads were corrected using the *correct* parameters in Canu with default nanopore settings (Koren et al., 2017). Canu requires information on the expected genome size so that coverage of the input reads can be determined; we used the size of the spinach chloroplast reference genome as the expected size in the whole genome approach (Experiment 4) and the expected target sequence length between gRNAs in all other experiments. The resulting sequences were then assembled de novo using Flye v2.9 (Kolmogorov et al., 2019). For mixed-species samples (Experiments 5–6), we used the metagenome assembly mode in Flye (metaFlye) with the *meta* and the *nano-corr* parameters as appropriate for error-corrected nanopore reads. To determine the percent identity between each contig and the reference genome, we mapped contigs to the reference genome(s) of target species(s) using minimap2 (Li, 2018) with the Oxford Nanopore option in Geneious Prime. When mapping contigs from the 16S rRNA (Experiment 3) and whole genome (Experiment 4) experiments to the spinach chloroplast, we enabled secondary alignments to allow reads to be mapped to multiple locations within the inverted repeats.

2.6 | Comparison of CRISPR-Cas enrichment and PCR-based DNA metabarcoding

We compared the hypothetical DNA sequence relative read abundance of target plant species in the mixed-species sample with empirical data obtained using both CRISPR-Cas and PCR-based sequencing approaches (Appendix S1). For the PCR-based benchmark, we used 2 × 150 bp Illumina sequencing and required a strict 100% identity between the resulting amplicon sequences and a global reference library (Appendix S1, Table S2). To calculate relative read abundance using sequences obtained with CRISPR-Cas enrichment, we used the read-count coverage of the contig built using Flye that mapped to the correct location in the reference genome of each target taxon. To calculate relative read abundance from DNA metabarcoding for the mixed-species sample, we converted sequence counts into proportional data. The relative read abundance values resulting from both methods can be interpreted as estimates of the proportional representation of DNA from the target taxa in the sample after accounting for all sources of bias and error, including variation in the tissue content of DNA per unit biomass, tissue homogenization and extraction, amplification and enrichment, sequencing accuracy, and bioinformatic processes.

3 | RESULTS

3.1 | Coverage of guide RNAs (gRNAs)

In total, we identified 54,837 unique gRNA sequences across the reference chloroplast genomes of the 7 target species (9,852–10,817 unique gRNAs per species; all identified gRNAs can be found in Table S1). This included (*i*) 44,510 'narrow-coverage' gRNAs that were present in only one target species, (*ii*) 398 'broad-coverage' gRNAs that occurred across all 7 target species, (*iii*) 44,647 'high-fidelity' gRNAs that had only a single cut site in ≥1 target species and (*iv*) 10,190 'low-fidelity' gRNAs that matched multiple sites in ≥1 target species. Of the 44,647 high-fidelity gRNAs, 6,581 perfectly matched (100% identity) the reference chloroplast of at least 2 target species and 45 of these occurred in all 7 target species (Table S1 reports the set of target species' genomes that included each gRNA). Of the 10,190 low-fidelity gRNAs, 3,746 perfectly matched the reference genome for at least 2 target species and 353 matched all 7 species (Table S1). Due to structural differences in the chloroplast genomes of Poaceae, many gRNAs that had perfect homology and a single cut site in the reference genomes of wheat, oat, and corn did not appear in spinach, sunflower, soy, or peanut. Nevertheless, we identified many potential broad-coverage gRNAs, especially when allowing for ≤2 bp mismatches at the 5′ end of the gRNAs. Moreover, 14 of these broad-coverage gRNAs were located within 3 kb of *rbc*L, 16 were located within 3 kb of *mat*K, and 2 were located within 3 kb *trn*L-P6 across all 7 target species. These CRISPR-associated loci present opportunities to develop broad-spectrum enrichment protocols for DNA barcoding and metabarcoding studies.

3.2 | Barcoding approach

We obtained high coverage and accuracy sequencing multiple plant DNA barcodes (Experiments 1–3). Experiment 1 targeted *rbc*L using one forward and one reverse gRNA and yielded a total of 1,531 reads with a *Q*-score ≥8 (Table 2, Figures 2 and 3). Raw sequence lengths of 130–15,971 bp (mean: 4004 bp) encompassed the target length of 5809 bp. Reads were then error-corrected using Canu and 44 consensus reads were produced; a subset of 41 (93%) reads mapped to the reference spinach chloroplast and 35 mapped to *rbc*L (85% of mapped corrected reads; Table 2). De novo assembly using Flye generated 2 contigs with lengths of 7,048 and 5,755 bp that

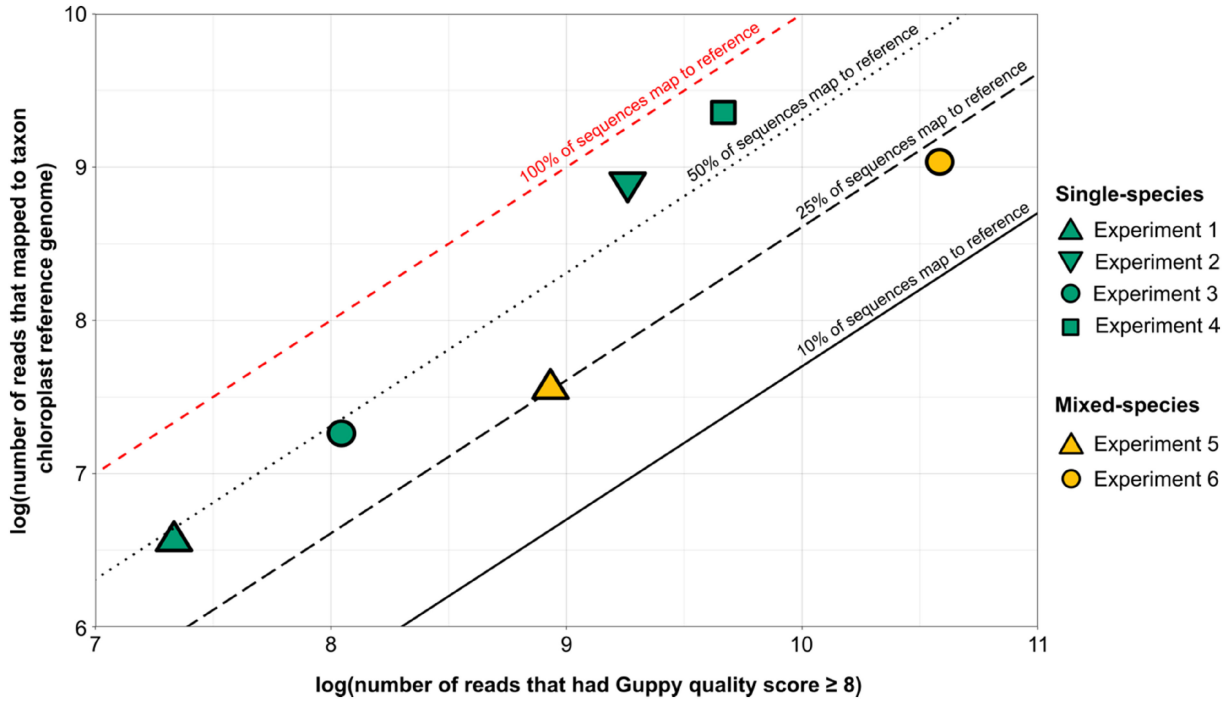

FIGURE 2 Comparison of the sequence reads generated and successfully mapped to reference genomes in each experiment. The number of DNA sequence reads that passed the Guppy basecaller ($Q \geq 8$) is shown on the *x*-axis and the number of those reads that mapped to the appropriate chloroplast reference genomes is shown on the *y*-axis. The four single-species experiments had a greater proportion of base-called reads that mapped to the reference chloroplast genome of the target taxon compared to mixed-species experiments which included six target taxa.

both mapped to the spinach reference, but only the shorter contig aligned to the target region (31× coverage with 99.5% pairwise identity; Table 2). The longer contig did not align to the target region but to an upstream region of the chloroplast (8× coverage).

Experiment 2 targeted multiple plant DNA barcodes with overlapping forward and reverse gRNAs, yielding a total of 10,506 reads with a Q-score ≥8 (Table 2, Figures 2 and 3). Raw sequence lengths ranged from 126 bp to 19,890 bp (mean: 4149 bp); when mapped to the reference spinach chloroplast, most raw reads sat downstream of the forward gRNA (Figure 3b) which was unexpected given that two gRNAs were used that ran in opposite directions from a single enrichment site. In total, 38 error-corrected reads were produced; 33 (87%) of these corrected consensus reads mapped to the spinach chloroplast reference genome. Of the corrected consensus reads that mapped to the spinach chloroplast reference genome, 32 reads aligned downstream of the forward gRNA and 1 read overlapped the forward and reverse gRNA (Table 2). De novo assembly generated 1 contig of 10,062 bp that aligned to the target region in the spinach chloroplast reference genome at 32× coverage and 99.2% identity to the reference (Table 2); however, the assembled contig sat downstream of the forward gRNA and therefore only included one (*atp*F-*atp*H) of the four target barcodes (*mat*K, *trn*H-*psb*A and *psb*K-*psb*I not included).

Experiment 3 targeted the 16S rRNA inverted repeat region of the chloroplast using a single gRNA. A total of 3,117 reads had a Q-score ≥8 with a mean sequence length of 4,023 bp (range: 139–22,583 bp) which span the length of the 16S rRNA region (Table 2, Figures 2 and 3). A total of 54 error-corrected reads were produced; 34 (63%) of these corrected consensus reads were mapped to the spinach chloroplast reference and all aligned to the target region (Table 2). De novo assembly generated 1 contig of 8,911 bp that mapped to the correct two locations within the chloroplast genome with 31× coverage and 99.4% identity to the reference sequence (Table 2).

3.3 | Whole chloroplast approach

The CRISPR-based enrichment approach yielded high sequencing depth of coverage and accuracy in sequencing the spinach chloroplast genome (Experiment 4). We obtained 15,766 reads with a Q-score ≥8 and a mean read length of 4328 bp (range: 109–25,372 bp; Table 2, Figures 2 and 3). A total of 1256 error-corrected reads were produced and 1118 (89%) of these mapped to the spinach chloroplast reference genome (Table 2). De novo assembly generated 9 contigs of 1733–18,510 bp that all aligned to the reference genome (Table 2). Of the 9 contigs, 2 contigs occurred in the inverted repeat regions. Together, the 9 contigs covered 81% of the spinach chloroplast reference genome (121,284 bp of 150,725 bp) with an average 60× coverage (20×–128× coverage across contigs) and provided excellent accuracy with 99.3%–99.6% identity to the reference genome (Table 2).

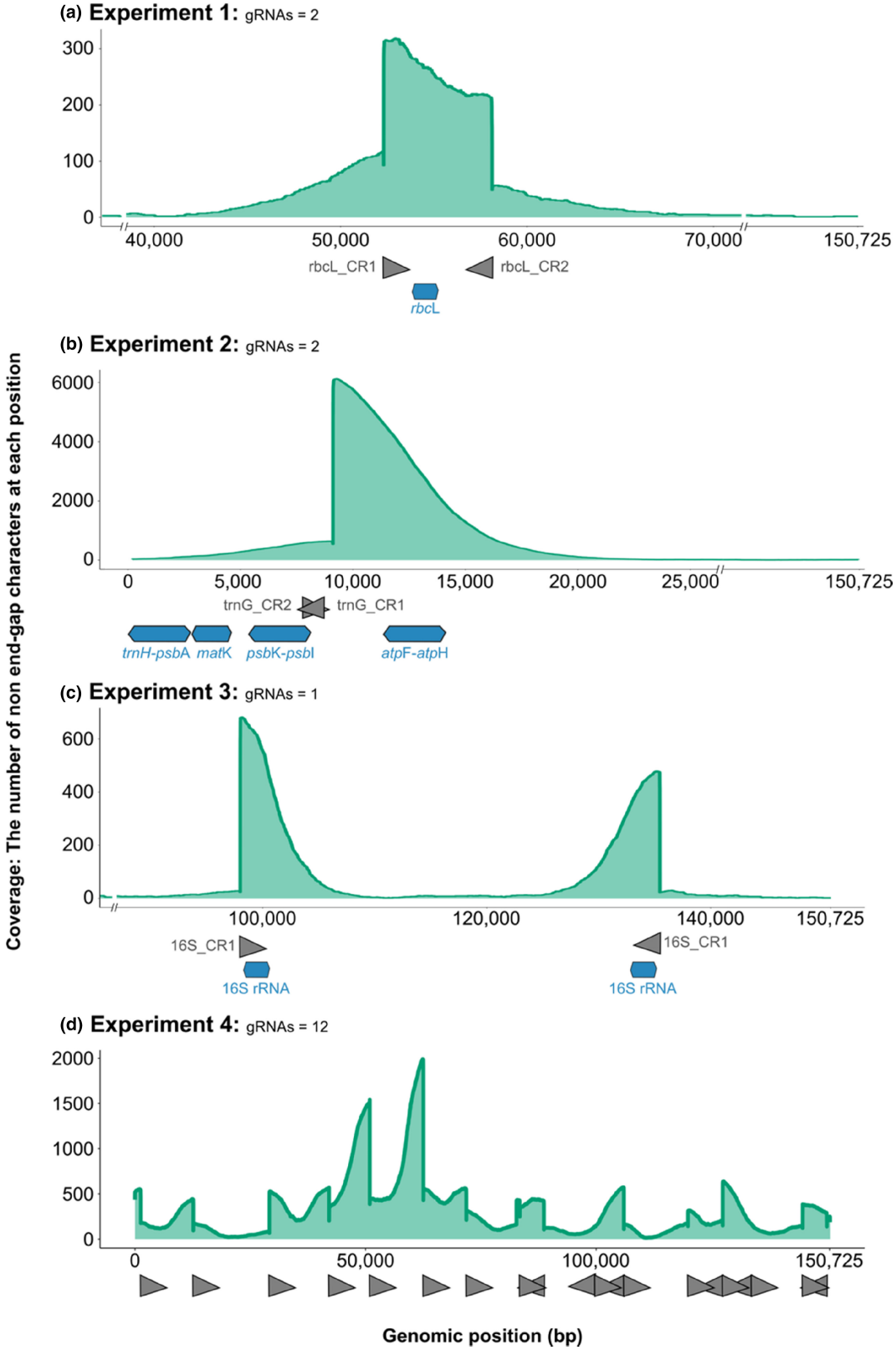

FIGURE 3 Coverage of raw sequence reads that passed Guppy basecalling (Q ≥ 8) and were mapped to the spinach reference genome. In experiments 1–4, we enriched for (a) the *rbc*L plant barcode region, (b) multiple standard plant barcodes including *mat*K and *trn*H-*psb*A, (c) the 16S rRNA inverted repeat regions, and (d) the whole chloroplast genome. In panels (a) to (c), blue hexagons indicate the positions of target barcodes in the spinach reference chloroplast genome. In all panels, grey arrowheads identify the gRNAs binding sites (Table 1). Gaps in non-target sections of the reference genome are shown using –//– notation.

3.4 | Mixed-species approach

In vitro, we had varied depths of sequence coverage and accuracy in sequencing the *rbc*L barcode from a mixed set of 6 target species: soy, wheat, corn, peanut, sunflower, and oats (Experiments 5–6). For Experiment 5, 2 gRNAs were used to target the *rbc*L gene, whereas we used only 1 gRNA in Experiment 6 (Figure 1). Compared to Experiment 5, we obtained 5.2-fold more reads that had a Q-score ≥8 in Experiment 6 (Figure 2). When using metaFlye to generate de novo contigs, a low number of contigs were produced which meant that we inevitably failed to recover the full taxonomic breadth of the six species included in the samples (Table 2). We therefore tried a second approach to contig assembly where trimmed reads were corrected and assembled into contigs independently for each target taxon (using Flye). For both Experiments 5 and 6, we found that building contigs independently for each target taxon resulted in better pairwise identity, taxonomic breadth, and contig coverage (Table 2). On average, these contigs mapped the reference chloroplast genomes of the 6 target taxa with greater percent identity in Experiment 5, but with greater average coverage in Experiment 6 (Table 2); thus, when comparing methods that employed 2 gRNAs (Experiment 5) versus 1 gRNA (Experiment 6) to target the *rbc*L gene in a mixed-species sample, we obtained better accuracy with 2 gRNAs but better depth of coverage with 1 gRNA (Table 2). When focusing our analyses on the standard *rbc*L barcode region (553 bp), which was contained within the much longer contigs we generated, the 'barcode region' yielded better accuracy than the 'non-barcode region' in both Experiment 5 (86.9%–99.5%) and Experiment 6 (86.7%–100%).

Given that Experiment 6 generated greater contig coverage (Table 2, Figure 2), we investigated where the raw sequence reads that passed Guppy basecalling (Q ≥ 8) mapped to on the reference genomes of the target taxa in the mixed-species sample (Figure 4). We found some striking patterns when visualizing raw read coverage in this mixed-species sample. First, peak coverage differed in location for each of the target taxon and hence indicated the chloroplast rearrangements that have occurred across the angiosperms (Figure 4). Second, the gRNA used in Experiment 6 had a forward directionality in all target-taxa, except peanut and soybean where it had a reverse directionality. Finally, corn, wheat and oat showed a double-peaked raw read coverage 'topology' that differed from that of the other three target taxon which only showed a single peak in coverage. This double-peaked coverage pattern likely occurs because of a low number of raw reads that included nucleotides that are not present in the reference genomes of these taxa.

Our final goal for the mixed-species approach was to compare CRISPR- and PCR-based methods for estimating DNA sequence relative read abundance. Despite differences in the number of gRNAs used and data yield for Experiments 5 and 6, both experiments produced contigs corresponding to all 6 taxa in relatively even proportions compared to PCR, showing a greater resemblance to the a priori expectations based on biomass (Figure 5). Specifically, the CRISPR-based strategies yielded more accurate estimates of DNA relative read abundances for the three taxa that were of known equal (1/6th) biomass proportions (Experiment 5: oat, 19%; sunflower, 15%; peanut, 13%; Experiment 6: oat, 22%; sunflower, 19%; peanut, 12%) compared to PCR (oat, 1%; sunflower, 35%; peanut, 38%; Figure 5, Table S2).

When classifying sequences to the three major plant lineages included in the mixed-species sample (i.e., monocots, superrosids, and superasterids), Experiments 5 and 6 also produced estimates of DNA relative read abundances that were more similar to a priori proportional expectations than PCR (Figure S1).

4 | DISCUSSION

Although CRISPR is generally underutilized in the environmental sciences (Phelps et al., 2020), CRISPR-based enrichment strategies have shown promise and versatility (Baerwald et al., 2023; López-Girona et al., 2020; Ramón-Laca et al., 2023; Sánchez et al., 2022; Sandoval-Quintana et al., 2023; Stangl et al., 2020; Williams et al., 2023). We evaluated strategies to harness this power for important applications in environmental biology such as overcoming the plant DNA barcode resolution problem (CBOL Plant Working Group, 2009; Kress, 2017) and issues with PCR-based DNA relative read abundance calculations (Deagle et al., 2019; Littleford-Colquhoun, Freeman, et al., 2022). Here, we were able to (*i*) identify many broad-spectrum gRNAs within the chloroplast genomes of phylogenetically disparate and economically important taxa, (*ii*) demonstrate methodological versatility for sequencing plant DNA barcode loci, (*iii*) enrich and assemble a nearly complete chloroplast genome using just 12 gRNAs, and (*iv*) profile plant DNA within a mixed sample with evidence for both accuracy and precision.

This study demonstrated the versatility of CRISPR-based enrichment approaches that extend beyond taxon-specific detection to include 'universal' methods that work across a broad swath of the plant phylogeny. Initially, we identified a total of 54,837 candidate gRNAs across 7 target angiosperm species; 81% of these gRNAs occurred in only one target species and 19% occurred in ≥2 (Table S1), highlighting the possibility of achieving both taxon-specific and broad-range detection and sequencing of economically and ecologically relevant species. As more plant reference genomes become available (e.g., on GenBank), the ability to accurately identify targets that provide either broad- or narrow-spectrum coverage across taxa will only improve. In addition to validating several methods to enrich plant DNA barcodes from a sample containing a single species (Experiments 1–3), we also succeeded in multiplexing gRNAs to sequence most of the spinach chloroplast genome (Experiment 4). Methods for whole genome assembly using CRISPR enrichment can be benchmarked against related methods such as targeted probe sets (Johnson et al., 2019). Perhaps most promisingly, *rbc*L barcode sequences we obtained from mixed-species samples using CRISPR enrichment (Experiments 5–6) provided longer sequences and more accurate representation of relative abundances compared to the expected species biomasses than a widely used PCR-based method for

TABLE 2 Results from each step in bioinformatic pipeline.

Experiment	Target plant species	RefSeq accession of reference genome for the target plant species	Number of base-called reads (guppy)	Number of Q ≥ 8 reads (guppy)[a]	Number of trimmed reads (Porechop)	Number of consensus (error) corrected reads (Canu)
Barcoding approach						
Experiment 1	Spinach	NC_002202	2540	1531	1531	44
Experiment 2	Spinach	NC_002202	12,700	10,506	10,506	38
Experiment 3	Spinach	NC_002202	4010	3117	1439	54
Whole chloroplast approach						
Experiment 4	Spinach	NC_002202	19,230	15,766	11,575	1256
Mixed-species approach						
2× gRNAs (metaFlye)						
Experiment 5	Corn	NC_001666	16,720	7569	7567	22
Experiment 5	Wheat	NC_002762	16,720	7569	7567	22
Experiment 5	Soybean	NC_007942	16,720	7569	7567	22
Experiment 5	Oat	NC_027468	16,720	7569	7567	22
Experiment 5	Peanut	NC_037358	16,720	7569	7567	22
Experiment 5	Sunflower	NC_007977	16,720	7569	7567	22
2× gRNAs (Flye)						
Experiment 5	Corn	NC_001666	16,720	7569	7567	102
Experiment 5	Wheat	NC_002762	16,720	7569	7567	98
Experiment 5	Soybean	NC_007942	16,720	7569	7567	88
Experiment 5	Oat	NC_027468	16,720	7569	7567	92
Experiment 5	Peanut	NC_037358	16,720	7569	7567	63
Experiment 5	Sunflower	NC_007977	16,720	7569	7567	73
1× gRNA (metaFlye)						
Experiment 6	Corn	NC_001666	52,830	39,507	39,491	8
Experiment 6	Wheat	NC_002762	52,830	39,507	39,491	8
Experiment 6	Soybean	NC_007942	52,830	39,507	39,491	8
Experiment 6	Oat	NC_027468	52,830	39,507	39,491	8
Experiment 6	Peanut	NC_037358	52,830	39,507	39,491	8
Experiment 6	Sunflower	NC_007977	52,830	39,507	39,491	8
1× gRNA (Flye)						
Experiment 6	Corn	NC_001666	52,830	39,507	39,491	149
Experiment 6	Wheat	NC_002762	52,830	39,507	39,491	147
Experiment 6	Soybean	NC_007942	52,830	39,507	39,491	156
Experiment 6	Oat	NC_027468	52,830	39,507	39,491	157
Experiment 6	Peanut	NC_037358	52,830	39,507	39,491	179
Experiment 6	Sunflower	NC_007977	52,830	39,507	39,491	115

Note: For experiments 1–6, we provide the target plant species and the accession number of the corresponding reference genome used to align output reads, the total number of Guppy base-called reads, the number of reads that passed Guppy quality control, the number of reads retained following adapter trimming in Porechop, the number of consensus (error) corrected reads produced using Canu, the number of corrected reads that mapped to the region of interest (on-target), the number of contigs assembled from these corrected reads using Flye/metaFlye, the number of assembled contigs that mapped to the target species reference genome, the number of assembled contigs mapped to the region of interest (on-target) and the mean fold-coverage, mean pairwise identity, and the length of all on-target contigs. Outputs for different approaches (barcoding, whole chloroplast and mixed species) are shown using dark grey banners. For the mixed-species approach (Experiments 5–6), we report separate outputs (light grey banners) depending on whether contigs were assembled using all corrected reads of a sequencing run (using metaFlye) or whether contigs were assembled independently for each target taxon (using Flye).

[a]Number of Guppy base-called reads that had a quality score ≥8.

[b]Number of corrected reads that mapped to target-taxon reference genome.

[c]Number of contigs that mapped to target-taxon reference genome.

[d]Number of on-target contigs that mapped to target-taxon reference genome.

[e]Mean pairwise identity between on-target contigs and target-taxon reference genome.

Number of on-target corrected reads[b]	Number of contigs built (Flye / metaFlye)	Number of mapped contigs[c]	Number of on-target contigs[d]	Mean coverage of on-target contigs	Mean pairwise identity of on-target contigs (%)[e]	Mean bp of on-target contigs
41	2	2	1	31×	99.5	5755
33	1	1	1	32×	99.2	10,062
34	1	1	1	31×	99.4	8911
1118	9	9	9	60×	99.5	11,367
15	2	2	1	6×	80.8	2827
15	2	2	1	6×	80.7	2827
19	2	2	2	6×	82.9	2827
14	2	2	1	6×	80.8	2827
19	2	2	2	6×	86.9	2827
18	2	2	2	6×	86.4	2827
101	2	2	2	37×	82.4	2961
98	2	2	2	37×	80.3	2530
88	2	2	2	28×	79.2	4140
89	2	2	2	37×	81.0	3840
63	2	2	1	25×	98.6	5992
73	2	2	2	29×	90.2	4528
4	1	1	1	7×	73.6	2968
3	1	1	1	7×	73.0	2968
8	1	1	1	7×	85.5	2968
4	1	1	1	7×	73.1	2968
8	1	1	1	7×	81.6	2968
8	1	1	1	7×	98.9	2968
149	3	3	2	39×	79.3	3183
146	2	2	1	66×	74.3	3604
156	2	2	2	28×	83.0	4702
157	2	2	1	60×	73.2	4167
179	2	2	2	32×	88.3	4104
115	1	1	1	52×	99.4	6077

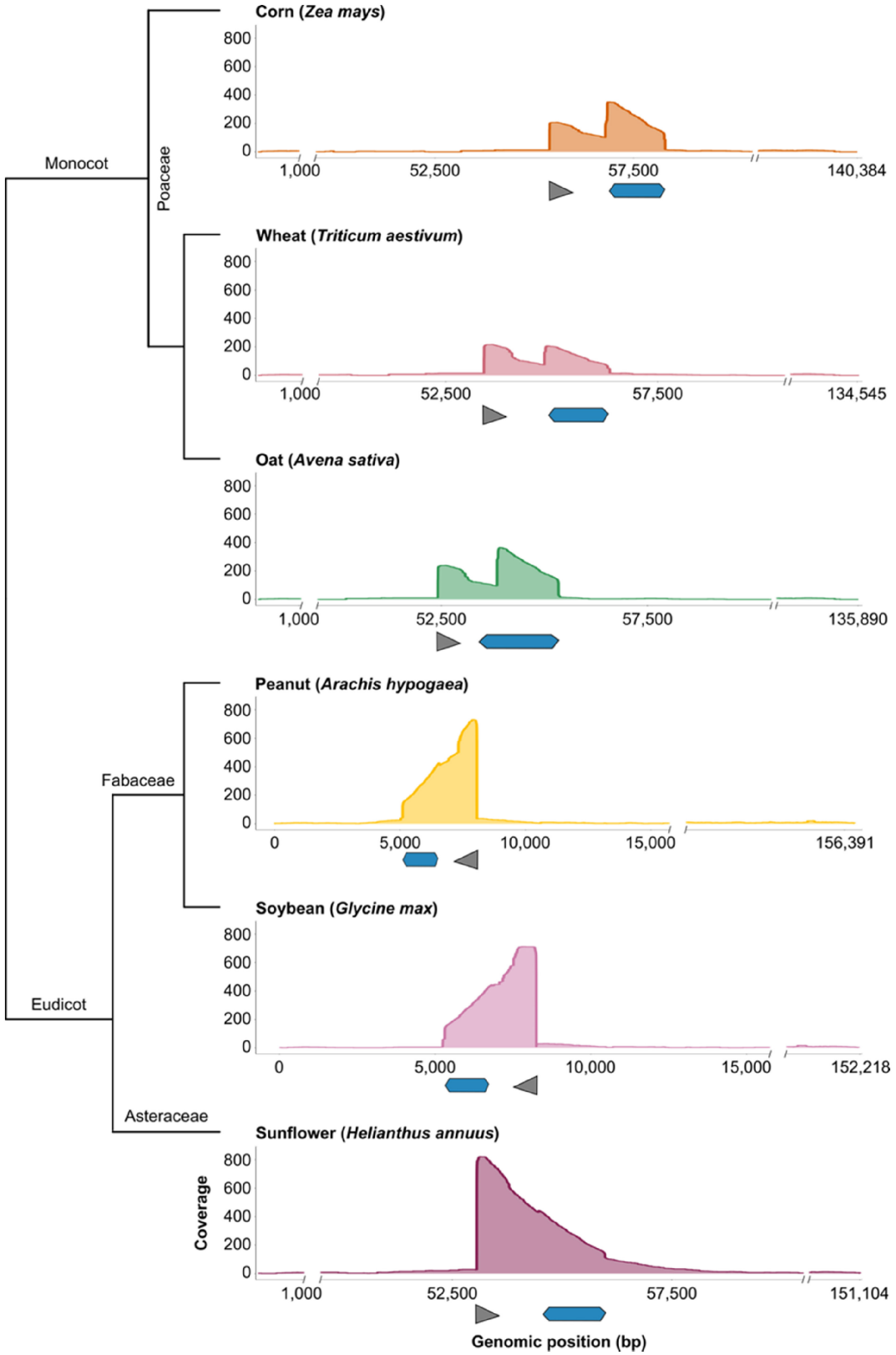

FIGURE 4 Depth of coverage for raw sequence reads that passed Guppy basecalling (Q ≥ 8) and mapped to reference genomes in mixed-species Experiment 6 using a single gRNA. The phylogeny of the six target taxa is shown to the left. Blue hexagons indicate the position of the *rbc*L gene, grey arrowheads indicate where the gRNA binds in each target-taxa, and the coverage values represent the number of non-end-gap characters obtained from sequences mapping to each position. Gaps in the off-target portions of the reference genomes are indicated using –//– notation. Peak coverage differs in location in each target taxon due to different chromosomal arrangements across taxa. The gRNA used (*rbc*L_CR1) was predicted to bind ~1500 bp up or downstream from *rbc*L in each reference genome.

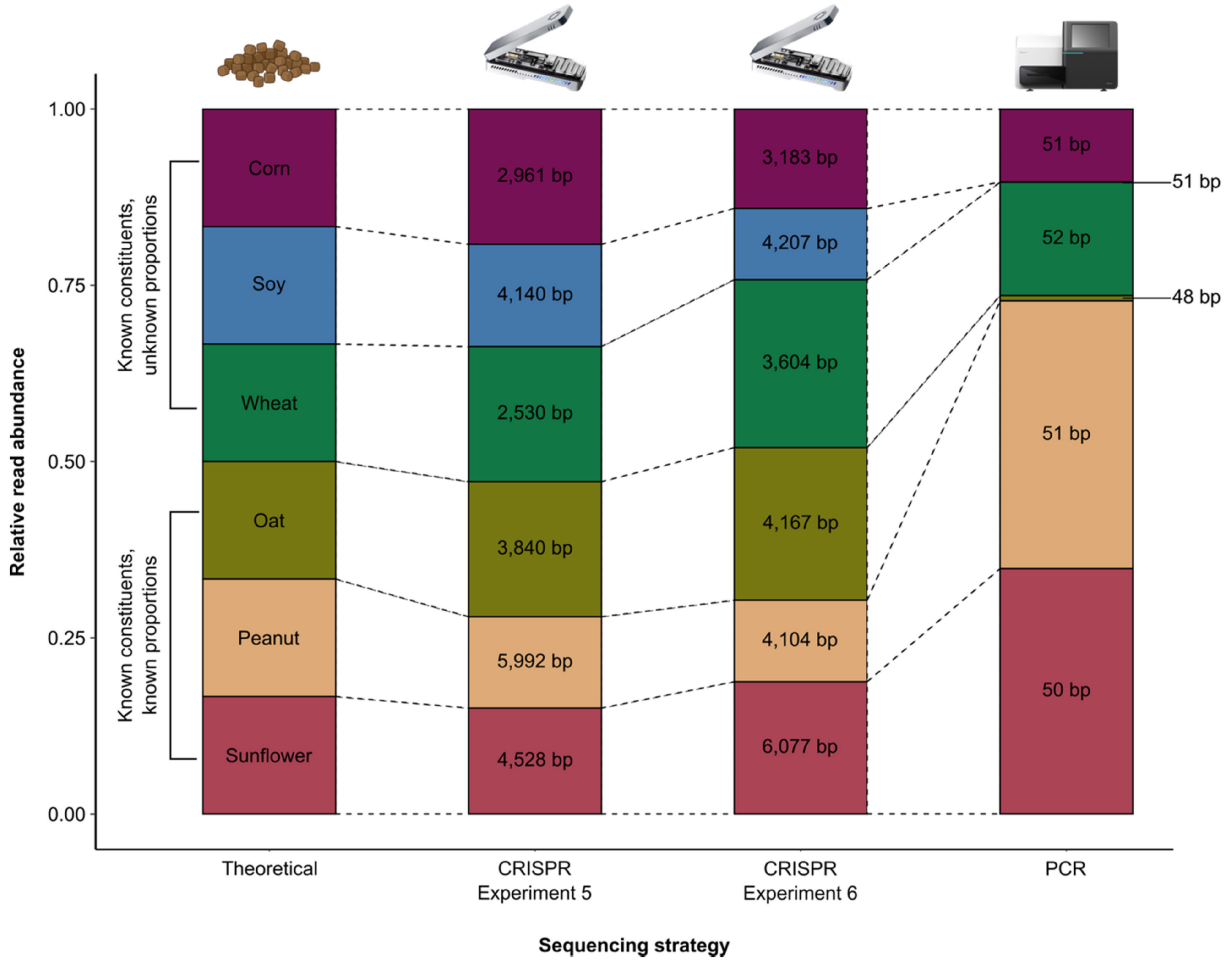

FIGURE 5 Stacked barplots comparing the results of CRIPSR- and PCR-based methods. From left to right, we show that the hypothetical relative biomass of each target taxa in the mixed-species sample, contig coverage for Experiment 5 of CRISPR-Cas enrichment (2× gRNA), contig coverage for Experiment 6 of CRISPR-Cas enrichment (1× gRNA), and relative read abundance obtained using PCR-based DNA metabarcoding. Of 6 taxa in the mixed sample, 3 were of known biomass proportions (oat, peanut and sunflower; ⅙ each) and 3 were of unknown biomass proportions (corn, soy, wheat). The contig and amplicon lengths generated per taxon for each experiment are shown in each segment of the barplot. All target taxa were detected with CRISPR-Cas enrichment which produced contig lengths per taxon that were at least 48-fold longer than amplicon sequencing.

DNA metabarcoding. Although the DNA content of plant cells going into the mixture and the final DNA concentration of the extracts could diverge, generating errant estimates of relative abundances based on sequence data, we found that two different CRISPR-nanopore protocols provided far closer matches to the expected proportions than PCR (Figure 5); results thus provide compelling evidence that the strategy is generalizable and that gRNAs may be interchangeable in efforts to obtain accurate and reproducible estimates of relative abundance. It is widely acknowledged that PCR-based methods can result in stochastic and biased abundance data (Pawluczyk et al., 2015), with strategies used to process such errors remaining largely contentious within the field (Littleford-Colquhoun, Freeman, et al., 2022; Littleford-Colquhoun, Sackett, et al., 2022). Thus, if CRISPR-Cas enrichment is capable of producing more accurate relative read abundance data and if we work collectively towards improving the strategy, then it could alleviate substantial consternation within the field. An important step towards this goal will be establishing general expectations about the efficacy of different mixed-species approaches (e.g., assays that utilize one gRNA vs. a multiplex of two or more).

Some of our CRISPR-Cas experiments yielded unexpected results that reveal opportunities to address future questions about the efficacy of different approaches. First, we encountered off-target enrichment in all experiments, but off-target activity was especially high in the mixed-species data. The samples used in mixed-species experiments were generally more highly processed and thus the templates were both more genetically diverse and potentially degraded (Experiments 5–6). Off-target activity suggests gRNAs may engage in some non-specific binding and/or encountered some structurally similar loci across the multiple genomes included in the sample (e.g., the nuclear genome or mitochondrial genome). Past work has shown that gRNAs can randomly bind to non-target regions, with Cas9 known to sometimes bind to non-canonical PAM sites (Kleinstiver et al., 2015). Such binding

may have contributed to non-target enrichment in Experiment 1, where we found a de novo contig that mapped upstream of the target region and in Experiments 5 and 6 where non-chloroplast contigs were generated. While many previous studies have shown some degree of off-target enrichment when using single-species samples (López-Girona et al., 2020; Ramón-Laca et al., 2023), Sandoval-Quintana et al. (2023) found that only 0.03% of good quality reads covered the region they wished to study when enriching a bacterial gene from a complex microbial sample, indicating that per-species coverage presents a challenge that may be amplified in more complex samples that include a greater diversity of both target and non-target DNA (e.g., Experiments 5 and 6). To address this challenge, careful gRNA design should be a priority and it may be instructive to experiment with mismatch tolerance values; by allowing ≤2 bp mismatch tolerance at the 5′ end of the gRNAs we were able to expand the number of candidate gRNAs for testing across the seven species used in these trials, but this type of decision may result in the selection of gRNAs that show more off-target activity than would be expected for a species-specific target. Computational methods that facilitate upstream screening of gRNAs for off-target activity along chromosomes of multiple taxa could help overcome these downstream challenges.

Perhaps the greatest need for further research required to translate CRISPR-based enrichment methods for the sequencing of complex mixtures will revolve around methodologies to build contigs. Due to the relatively low depth of coverage and percentage of base-called reads that mapped back to the reference genomes in Experiments 5–6, we were unable to construct species-specific contigs using metaFlye on the full dataset; we had to build contigs independently for each species using the raw reads that mapped back to each plant species' reference genome. In many real-world applications involving environmental DNA, there will not be *a priori* knowledge of reference genomes for all species in the mixture (Yang et al., 2021) and thus more sensitive taxon-calling methods will be required. A promising strategy involves translating bioinformatic methods that are being developed specifically for the analysis of bacterial metagenome-assembled genomes ('MAGs') for future applications involving CRISPR-based enrichment sequencing (Parks et al., 2017; Stewart et al., 2019; Tully et al., 2018), but our results suggest that achieving acceptable levels of accuracy may ultimately require more than simply 'tuning' the parameters used in these existing methods (e.g., Experiments 5–6).

Considerations for future experimental design at the bench could help enhance the versatility and accuracy of CRISPR-based sequencing methods for biodiversity research, especially for challenging mixed-species analyses. For example, DNA extraction methods (Kang et al., 2023; Russo et al., 2022), the number of purification steps used during library preparation (De La Cerda et al., 2023), the specific Cas system deployed (e.g., Cas9 vs. Cas3 or Cas12a [Schultzhaus et al., 2021]), and the sequencing platform utilized (e.g., Oxford Nanopore vs. Illumina or PacBio [Li & Harkess, 2018]) may need to be optimized in order to ensure adequate on-target sequence coverage. There are encouraging strategies to deplete non-target sequences, such as in host DNA in microbiome studies, using CRISPR-Cas selective amplicon sequencing (Zhong et al., 2021). Strategies to enhance the enrichment of on-target reads also include methods for tiling gRNAs, whereby overlapping gRNAs can be used to extend the enrichment of the target region (López-Girona et al., 2020), or to improve the median depth of coverage for a particular locus (Gilpatrick et al., 2020). In Experiment 2, however, we had mixed success with a tiling approach because only one of the tiled gRNAs was effective, leaving gaps across multiple barcode genes. Possible causes for this type of skewed enrichment include inadequate separation of gRNA target sites along the chromosome, use of gRNAs that run in different directions, preferential binding by one of the gRNAs, and/or ligation bias across CRISPR-Cas cut sites.

Our in silico analysis of plant gRNAs coupled with our six validation experiments provide proof of concept involving the use of CRISPR-based enrichment sequencing for use in environmental biology. This technology can be used to build accurate plant DNA barcode libraries with sequences that are long enough to span multiple barcode regions and thus overcome long-standing limitations to taxonomic resolution in PCR-based barcoding studies (CBOL Plant Working Group, 2009; Kress, 2017)—potentially providing sequences for entire chloroplast genomes—though overcoming the challenge of translating this potential into versatile and cost-effective methods for analysis of environmental DNA represents an exciting area for development (Schultzhaus et al., 2021). Moving forward, these approaches can be extended to incorporate other facets of research that are integral for biodiversity discovery, such as determining structural rearrangements (Li et al., 2016; Ramón-Laca et al., 2023; Sun et al., 2022), phylogenetic patterns (Yang et al., 2014), targeted genome sequencing (López-Girona et al., 2020), multiplexing samples and loci within a single reaction (Stangl et al., 2020; Welch et al., 2022), and building metagenome-assembled genomes (Liu et al., 2022; Sandoval-Quintana et al., 2023).

ACKNOWLEDGEMENTS

We would like to thank Colin Donihue, Alex Harkess, and Timothy Divoll. Funding was provided by IBES seed funding at Brown University and NSF DEB-2046797.

CONFLICT OF INTEREST

Authors declare that there are no conflicts of interest.

DATA AVAILABILITY STATEMENT

Nanopore sequence read data and sample metadata for all experiments have been made available at NCBI (BioProject: PRJNA989586). Bioinformatic steps taken for each experiment and code used for building the global plant reference library can be found on the Brown Digital Repository (https://doi.org/10.26300/m1kr-3q85). Illumina sequence read data and sample metadata for the mixed-species sample have been made available at NCBI (BioProject: PRJNA989255) (Littleford-Colquhoun & Kartzinel, 2023a, 2023b, 2023c).

ORCID

Bethan Littleford-Colquhoun https://orcid.org/0000-0002-2594-0061
Tyler R. Kartzinel https://orcid.org/0000-0002-8488-0580

REFERENCES

Baerwald, M. R., Funk, E. C., Goodbla, A. M., Campbell, M. A., Thompson, T., Meek, M. H., & Schreier, A. D. (2023). Rapid CRISPR-Cas13a genetic identification enables new opportunities for listed Chinook salmon management. *Molecular Ecology Resources*, 1–13.

Broughton, J. P., Deng, X., Yu, G., Fasching, C. L., Servellita, V., Singh, J., Miao, X., Streithorst, J. A., Granados, A., & Sotomayor-Gonzalez, A. (2020). CRISPR–Cas12-based detection of SARS-CoV-2. *Nature Biotechnology*, *38*(7), 870–874.

Caporaso, J. G., Lauber, C. L., Walters, W. A., Berg-Lyons, D., Huntley, J., Fierer, N., Owens, S. M., Betley, J., Fraser, L., & Bauer, M. (2012). Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. *The ISME Journal*, *6*(8), 1621–1624.

CBOL Plant Working Group. (2009). A DNA barcode for land plants. *Proceedings of the National Academy of Sciences of the United States of America*, *106*, 12794–12797.

Clarke, L. J., Soubrier, J., Weyrich, L. S., & Cooper, A. (2014). Environmental metabarcodes for insects: In silico PCR reveals potential for taxonomic bias. *Molecular Ecology Resources*, *14*(6), 1160–1170.

Cui, Y., Xu, J., Cheng, M., Liao, X., & Peng, S. (2018). Review of CRISPR/Cas9 sgRNA design tools. *Interdisciplinary Sciences: Computational LIfe Sciences*, *10*, 455–465.

Cunningham, C. H., Hennelly, C. M., Lin, J. T., Ubalee, R., Boyce, R. M., Mulogo, E. M., Hathaway, N., Thwai, K. L., Phanzu, F., & Kalonji, A. (2021). A novel CRISPR-based malaria diagnostic capable of plasmodium detection, species differentiation, and drug-resistance genotyping. *eBioMedicine*, *68*, 103415.

De La Cerda, G. Y., Landis, J. B., Eifler, E., Hernandez, A. I., Li, F. W., Zhang, J., Tribble, C. M., Karimi, N., Chan, P., & Givnish, T. (2023). Balancing read length and sequencing depth: Optimizing Nanopore long-read sequencing for monocots with an emphasis on the Liliales. *Applications in Plant Sciences*, *11*(3), e11524.

Deagle, B. E., Thomas, A. C., McInnes, J. C., Clarke, L. J., Vesterinen, E. J., Clare, E. L., Kartzinel, T. R., & Eveson, J. P. (2019). Counting with DNA in metabarcoding studies: How should we convert sequence reads to dietary data? *Molecular Ecology*, *28*(2), 391–406.

Doench, J. G., Fusi, N., Sullender, M., Hegde, M., Vaimberg, E. W., Donovan, K. F., Smith, I., Tothova, Z., Wilen, C., & Orchard, R. (2016). Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. *Nature Biotechnology*, *34*(2), 184–191.

Fu, B. X., St. Onge, R. P., Fire, A. Z., & Smith, J. D. (2016). Distinct patterns of Cas9 mismatch tolerance in vitro and in vivo. *Nucleic Acids Research*, *44*(11), 5365–5377.

Gilpatrick, T., Lee, I., Graham, J. E., Raimondeau, E., Bowen, R., Heron, A., Downs, B., Sukumar, S., Sedlazeck, F. J., & Timp, W. (2020). Targeted nanopore sequencing with Cas9-guided adapter ligation. *Nature Biotechnology*, *38*(4), 433–438.

Gilpatrick, T., Wang, J. Z., Weiss, D., Norris, A. L., Eshleman, J. R., & Timp, W. (2023). IVT generation of guideRNAs for Cas9-enrichment Nanopore sequencing. *bioRxiv*. 2023.2002. 2007.527484.

Hebert, P., Bock, D., & Prosser, S. (2022). Interrogating 1000 insect genomes for NUMTs: A risk assessment for species scans. Authorea Preprints.

Hoban, M. L., Whitney, J., Collins, A. G., Meyer, C., Murphy, K. R., Reft, A. J., & Bemis, K. E. (2022). Skimming for barcodes: Rapid production of mitochondrial genome and nuclear ribosomal repeat reference markers through shallow shotgun sequencing. *PeerJ*, *10*, e13790.

Johnson, M. G., Pokorny, L., Dodsworth, S., Botigue, L. R., Cowan, R. S., Devault, A., Eiserhardt, W. L., Epitawalage, N., Forest, F., & Kim, J. T. (2019). A universal probe set for targeted sequencing of 353 nuclear genes from any flowering plant designed using k-medoids clustering. *Systematic Biology*, *68*(4), 594–606.

Kaminski, M. M., Abudayyeh, O. O., Gootenberg, J. S., Zhang, F., & Collins, J. J. (2021). CRISPR-based diagnostics. *Nature Biomedical Engineering*, *5*(7), 643–656.

Kang, M., Chanderbali, A., Lee, S., Soltis, D. E., Soltis, P. S., & Kim, S. (2023). High-molecular-weight DNA extraction for long-read sequencing of plant genomes: An optimization of standard methods. *Applications in Plant Sciences*, *11*(3), e11528.

Karlikow, M., Amalfitano, E., Yang, X., Doucet, J., Chapman, A., Mousavi, P. S., Homme, P., Sutyrina, P., Chan, W., & Lemak, S. (2023). CRISPR-induced DNA reorganization for multiplexed nucleic acid detection. *Nature Communications*, *14*(1), 1505.

Kartzinel, T. R., & Pringle, R. M. (2015). Molecular detection of invertebrate prey in vertebrate diets: Trophic ecology of Caribbean Island lizards. *Molecular Ecology Resources*, *15*(4), 903–914.

Ke, Y., Huang, S., Ghalandari, B., Li, S., Warden, A. R., Dang, J., Kang, L., Zhang, Y., Wang, Y., & Sun, Y. (2021). Hairpin-spacer crRNA-enhanced CRISPR/Cas13a system promotes the specificity of single nucleotide polymorphism (SNP) identification. *Advanced Science*, *8*(6), 2003611.

Keck, F., Couton, M., & Altermatt, F. (2022). Navigating the seven challenges of taxonomic reference databases in metabarcoding analyses. *Molecular Ecology Resources*, *23*, 742–755.

Kelly, R. P., Shelton, A. O., & Gallego, R. (2019). Understanding PCR processes to draw meaningful conclusions from environmental DNA studies. *Scientific Reports*, *9*(1), 1–14.

Kleinstiver, B. P., Prew, M. S., Tsai, S. Q., Topkar, V. V., Nguyen, N. T., Zheng, Z., Gonzales, A. P., Li, Z., Peterson, R. T., & Yeh, J.-R. J. (2015). Engineered CRISPR-Cas9 nucleases with altered PAM specificities. *Nature*, *523*(7561), 481–485.

Knott, G. J., & Doudna, J. A. (2018). CRISPR-Cas guides the future of genetic engineering. *Science*, *361*(6405), 866–869.

Kolmogorov, M., Yuan, J., Lin, Y., & Pevzner, P. A. (2019). Assembly of long, error-prone reads using repeat graphs. *Nature Biotechnology*, *37*(5), 540–546.

Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., Bergman, N. H., & Phillippy, A. M. (2017). Canu: Scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Research*, *27*(5), 722–736.

Kress, W. J. (2017). Plant DNA barcodes: Applications today and in the future. *Journal of Systematics and Evolution*, *55*(4), 291–307.

Li, F. W., & Harkess, A. (2018). A guide to sequence your favorite plant genomes. *Applications in Plant Sciences*, *6*(3), e1030.

Li, F.-W., Kuo, L.-Y., Pryer, K. M., & Rothfels, C. J. (2016). Genes translocated into the plastid inverted repeat show decelerated substitution rates and elevated GC content. *Genome Biology and Evolution*, *8*(8), 2452–2458.

Li, H. (2018). Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics*, *34*(18), 3094–3100.

Littleford-Colquhoun, B. L., Freeman, P. T., Sackett, V. I., Tulloss, C. V., McGarvey, L. M., Geremia, C., & Kartzinel, T. R. (2022). *The precautionary principle and dietary DNA metabarcoding: Commonly used abundance thresholds change ecological interpretation*. Wiley Online Library.

Littleford-Colquhoun, B. L., & Kartzinel, T. R. (2023a). *Nanopore sequence data*. SRA; Accession: PRJNA989586 [dataset].

Littleford-Colquhoun, B. L., & Kartzinel, T. R. (2023b). *Illumina sequence data*. SRA; Accession: PRJNA989255 [dataset].

Littleford-Colquhoun, B. L., & Kartzinel, T. R. (2023c). *CRISPR-Cas enrichment bioinfomatic pipeline*. Brown Digital Respository [dataset]. https://doi.org/10.26300/m1kr-3q85

Littleford-Colquhoun, B. L., Sackett, V. I., Tulloss, C. V., & Kartzinel, T. R. (2022). Evidence-based strategies to navigate complexity in

dietary DNA metabarcoding: A reply. *Molecular Ecology*, *31*(22), 5660–5665.

Liu, L., Yang, Y., Deng, Y., & Zhang, T. (2022). Nanopore long-read-only metagenomics enables complete and high-quality genome reconstruction from mock and complex metagenomes. *Microbiome*, *10*(1), 209.

López-Girona, E., Davy, M. W., Albert, N. W., Hilario, E., Smart, M. E., Kirk, C., Thomson, S. J., & Chagné, D. (2020). CRISPR-Cas9 enrichment and long read sequencing for fine mapping in plants. *Plant Methods*, *16*(1), 1–13.

Manhart, J. R. (1995). Chloroplast 16S rDNA sequences and phylogenetic relationships of fern allies and ferns. *American Fern Journal*, *85*, 182–192.

O'Donnell, J. L., Kelly, R. P., Lowell, N. C., & Port, J. A. (2016). Indexed PCR primers induce template-specific bias in large-scale DNA sequencing studies. *PLoS One*, *11*(3), e0148698.

Parks, D. H., Rinke, C., Chuvochina, M., Chaumeil, P.-A., Woodcroft, B. J., Evans, P. N., Hugenholtz, P., & Tyson, G. W. (2017). Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nature Microbiology*, *2*(11), 1533–1542.

Pawluczyk, M., Weiss, J., Links, M. G., Egaña Aranguren, M., Wilkinson, M. D., & Egea-Cortines, M. (2015). Quantitative evaluation of bias in PCR amplification and next-generation sequencing derived from metabarcoding samples. *Analytical and Bioanalytical Chemistry*, *407*, 1841–1848.

Phelps, M. P., Seeb, L. W., & Seeb, J. E. (2020). Transforming ecology and conservation biology through genome editing. *Conservation Biology*, *34*(1), 54–65.

Qiu, X.-Y., Zhu, L.-Y., Zhu, C.-S., Ma, J.-X., Hou, T., Wu, X.-M., Xie, S.-S., Min, L., Tan, D.-A., & Zhang, D.-Y. (2018). Highly effective and low-cost microRNA detection with CRISPR-Cas9. *ACS Synthetic Biology*, *7*(3), 807–813.

Quan, J., Langelier, C., Kuchta, A., Batson, J., Teyssier, N., Lyden, A., Caldera, S., McGeever, A., Dimitrov, B., & King, R. (2019). FLASH: A next-generation CRISPR diagnostic for multiplexed detection of antimicrobial resistance sequences. *Nucleic Acids Research*, *47*(14), e83.

Ramón-Laca, A., Gallego, R., & Nichols, K. M. (2023). Affordable de novo generation of fish mitogenomes using amplification-free enrichment of mitochondrial DNA and deep sequencing of long fragments. *Molecular Ecology Resources*, *23*, 818–832.

Russo, A., Mayjonade, B., Frei, D., Potente, G., Kellenberger, R. T., Frachon, L., Copetti, D., Studer, B., Frey, J. E., & Grossniklaus, U. (2022). Low-input high-molecular-weight DNA extraction for long-read sequencing from plants of diverse families. *Frontiers in Plant Science*, *13*, 1494.

Sánchez, E., Ali, Z., Islam, T., & Mahfouz, M. (2022). A CRISPR-based lateral flow assay for plant genotyping and pathogen diagnostics. *Plant Biotechnology Journal*, *20*(12), 2418–2429.

Sandoval-Quintana, E., Stangl, C., Huang, L., Renkens, I., Duran, R., van Haaften, G., Monroe, G., Lauga, B., & Cagnon, C. (2023). CRISPR-Cas9 enrichment, a new strategy in microbial metagenomics to investigate complex genomic regions: The case of an environmental integron. *Molecular Ecology Resources*, *23*, 1288–1298.

Schultzhaus, Z., Wang, Z., & Stenger, D. (2021). CRISPR-based enrichment strategies for targeted sequencing. *Biotechnology Advances*, *46*, 107672.

Shashank, P. R., Parker, B. M., Rananaware, s., Plotkin, D., Couch, C., Yang, L. G., Nguyen, L. T., Prasannakumar, N., Braswell, W. E., & Jain, P. K. (2023). CRISPR-based diagnostics detects invasive insect pests. *bioRxiv*, 2023.2005.2016.541004.

Sledzinski, P., Nowaczyk, M., & Olejniczak, M. (2020). Computational tools and resources supporting CRISPR-Cas experiments. *Cell*, *9*(5), 1288.

Srivathsan, A., Lee, L., Katoh, K., Hartop, E., Kutty, S. N., Wong, J., Yeo, D., & Meier, R. (2021). ONTbarcoder and MinION barcodes aid biodiversity discovery and identification by everyone, for everyone. *BMC Biology*, *19*(1), 1–21.

Stangl, C., de Blank, S., Renkens, I., Westera, L., Verbeek, T., Valle-Inclan, J. E., González, R. C., Henssen, A. G., van Roosmalen, M. J., & Stam, R. W. (2020). Partner independent fusion gene detection by multiplexed CRISPR-Cas9 enrichment and long read nanopore sequencing. *Nature Communications*, *11*(1), 2861.

Stapleton, T. E., Weinstein, S. B., Greenhalgh, R., & Dearing, M. D. (2022). Successes and limitations of quantitative diet metabarcoding in a small, herbivorous mammal. *Molecular Ecology Resources*, *22*(7), 2573–2586.

Stewart, R. D., Auffret, M. D., Warr, A., Walker, A. W., Roehe, R., & Watson, M. (2019). Compendium of 4,941 rumen metagenome-assembled genomes for rumen microbiome biology and enzyme discovery. *Nature Biotechnology*, *37*(8), 953–961.

Strauss, S. H., Palmer, J. D., Howe, G. T., & Doerksen, A. H. (1988). Chloroplast genomes of two conifers lack a large inverted repeat and are extensively rearranged. *Proceedings of the National Academy of Sciences of the United States of America*, *85*(11), 3898–3902.

Sun, M., Zhang, M., Chen, X., Liu, Y., Liu, B., Li, J., Wang, R., Zhao, K., & Wu, J. (2022). Rearrangement and domestication as drivers of Rosaceae mitogenome plasticity. *BMC Biology*, *20*(1), 181.

Tully, B. J., Graham, E. D., & Heidelberg, J. F. (2018). The reconstruction of 2,631 draft metagenome-assembled genomes from the global oceans. *Scientific Data*, *5*(1), 1–8.

Vences, M., Thomas, M., Van der Meijden, A., Chiari, Y., & Vieites, D. R. (2005). Comparative performance of the 16S rRNA gene in DNA barcoding of amphibians. *Frontiers in Zoology*, *2*(1), 1–12.

Wang, Y., Huang, C., & Zhao, W. (2022). Recent advances of the biological and biomedical applications of CRISPR/Cas systems. *Molecular Biology Reports*, *49*(7), 7087–7100.

Welch, N. L., Zhu, M., Hua, C., Weller, J., Mirhashemi, M. E., Nguyen, T. G., Mantena, S., Bauer, M. R., Shaw, B. M., & Ackerman, C. M. (2022). Multiplexed CRISPR-based microfluidic platform for clinical testing of respiratory viruses and identification of SARS-CoV-2 variants. *Nature Medicine*, *28*(5), 1083–1094.

Wick, R. R., Judd, L. M., Gorrie, C. L., & Holt, K. E. (2017). Completing bacterial genome assemblies with multiplex MinION sequencing. *Microbial Genomics*, *3*(10), e000132.

Williams, M. A., de Eyto, E., Caestecker, S., Regan, F., & Parle-McDermott, A. (2023). Development and field validation of RPA-CRISPR-Cas environmental DNA assays for the detection of brown trout (Salmo trutta) and Arctic char (*Salvelinus alpinus*). *Environmental DNA*, *5*(2), 240–250.

Williams, M. A., Hernandez, C., O'Sullivan, A. M., April, J., Regan, F., Bernatchez, L., & Parle-McDermott, A. (2021). Comparing CRISPR-Cas and qPCR eDNA assays for the detection of Atlantic salmon (*Salmo salar* L.). *Environmental DNA*, *3*(1), 297–304.

Williams, M. A., O'Grady, J., Ball, B., Carlsson, J., de Eyto, E., McGinnity, P., Jennings, E., Regan, F., & Parle-McDermott, A. (2019). The application of CRISPR-Cas for single species identification from environmental DNA. *Molecular Ecology Resources*, *19*(5), 1106–1114.

Wilson, L. O., O'Brien, A. R., & Bauer, D. C. (2018). The current state and future of CRISPR-Cas9 gRNA design tools. *Frontiers in Pharmacology*, *9*, 749.

Xie, K., Minkenberg, B., & Yang, Y. (2015). Boosting CRISPR/Cas9 multiplex editing capability with the endogenous tRNA-processing system. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(11), 3570–3575.

Xu, Y., & Li, Z. (2020). CRISPR-Cas systems: Overview, innovations and applications in human disease research and gene therapy. *Computational and Structural Biotechnology Journal*, *18*, 2401–2415.

Yang, C., Chowdhury, D., Zhang, Z., Cheung, W. K., Lu, A., Bian, Z., & Zhang, L. (2021). A review of computational tools for generating metagenome-assembled genomes from metagenomic sequencing data. *Computational and Structural Biotechnology Journal*, *19*, 6301–6314.

Yang, J. B., Li, D. Z., & Li, H. T. (2014). Highly effective sequencing whole chloroplast genomes of angiosperms by nine novel universal primer pairs. *Molecular Ecology Resources*, *14*(5), 1024–1031.

Zhong, K. X., Cho, A., Deeg, C. M., Chan, A. M., & Suttle, C. A. (2021). Revealing the composition of the eukaryotic microbiome of oyster spat by CRISPR-Cas selective amplicon sequencing (CCSAS). *Microbiome*, *9*, 1–17.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Littleford-Colquhoun, B., & Kartzinel, T. R. (2023). A CRISPR-based strategy for targeted sequencing in biodiversity science. *Molecular Ecology Resources*, *00*, e13920. https://doi.org/10.1111/1755-0998.13920