Received: 8 July 2021 | Accepted: 15 November 2021

DOI: 10.1111/2041-210X.13784

Methods in Ecology and Evolution BRITISH ECOLOGICAL SOCIETY

RESEARCH ARTICLE

ISSRseq: An extensible method for reduced representation sequencing

Brandon T. Sinn[1,2] | Sandra J. Simon[2,3,4] | Mathilda V. Santee[2] | Stephen P. DiFazio[2] | Nicole M. Fama[2,5] | Craig F. Barrett[2]

[1]Department of Biology and Earth Science, Otterbein University, Westerville, OH, USA

[2]Department of Biology, West Virginia University, Morgantown, WV, USA

[3]Institute for Sustainability, Energy, and Environment (ISEE), University of Illinois at Urbana-Champaign, Urbana, IL, USA

[4]Department of Biology, West Virginia University Institute of Technology, Beckley, WV, USA

[5]Genetic Immunotherapy Section, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, MD, USA

Correspondence
Brandon T. Sinn
Email: sinn1@otterbein.edu

Funding information
National Science Foundation, Grant/Award Number: 1542509 and 1920858; West Virginia Research Challenge Fund; West Virginia University; WVU Core Facility; CTSI, Grant/Award Number: #U54 GM104942; WVU Department of Biology; WVU Eberly College of Arts and Sciences; WVU Program to Stimulate Competitive Research

Handling Editor: Susan Johnston

Abstract

1. The capability to generate densely sampled single nucleotide polymorphism (SNP) data is essential in diverse subdisciplines of biology, including crop breeding, pathology, forensics, forestry, ecology, evolution and conservation. However, the wet-laboratory expertise and bioinformatics training required to conduct genome-scale variant discovery remain limiting factors for investigators with limited resources.
2. Here we present ISSRseq, a PCR-based method for reduced representation of genomic variation using simple sequence repeats as priming sites to sequence inter simple sequence repeat (ISSR) regions. Briefly, ISSR regions are amplified with single primers, pooled, used to construct sequencing libraries with a commercially available kit, and sequenced on the Illumina platform. We also present a flexible bioinformatic pipeline that assembles ISSR loci, calls and hard filters variants, outputs data matrices in common formats, and conducts population analyses using R.
3. Using three angiosperm species as case studies, we demonstrate that ISSRseq is highly repeatable, necessitates only simple wet-laboratory skills and commonplace instrumentation, is flexible in terms of the number of single primers used, and can generate genomic-scale variant discovery on par with existing RRS methods which require more complex wet-laboratory procedures.
4. ISSRseq represents a straightforward approach to SNP genotyping in any organism, and we predict that this method will be particularly useful for those studying population genomics and phylogeography of non-model organisms. Furthermore, the ease of ISSRseq relative to other RRS methods should prove useful to those lacking advanced expertise in wet-laboratory methods or bioinformatics.

KEYWORDS
Corallorhiza, ISSRseq, population genomics, reduced representation sequencing

Brandon T. Sinn and Sandra J. Simon contributed equally to this work.

1 | INTRODUCTION

Reduced representation sequencing methods (RRS), in concert with high-throughput sequencing technologies, have revolutionized research in agriculture, conservation, ecology, evolutionary biology, forestry, population genetics and systematics (Altshuler et al., 2000; Davey et al., 2011; Lemmon & Lemmon, 2013; Meek & Larson, 2019; Narum et al., 2013; Van Tassell et al., 2008). These methods generate broad surveys of genomic diversity by sampling only a fraction of the genome, therefore allowing for multiple samples to be sequenced simultaneously which reduces sequencing costs (Franchini et al., 2017; Peterson et al., 2012). While there exists a bewildering array of RRS methods (Campbell et al., 2018), the most commonly employed are various forms of restriction-site-associated DNA sequencing (RAD; Baird et al., 2008) and genotyping-by-sequencing (GBS; Elshire et al., 2011). Both involve fragmenting the genome with one or more restriction enzymes, ligating adapters and unique barcode sequences, pooling, optionally size-selecting the fragments for optimal sequencing length, amplifying resulting fragments and sequencing typically with Illumina short-read technology (Davey et al., 2011). Flexibility in these methods lies principally in the selection of one or more restriction enzymes; for example, one popular method, double-digest RAD sequencing, uses both a 'rare-cutting' and 'common-cutting' enzyme (Peterson et al., 2012).

While methods such as GBS and RAD are increasingly commonplace, they may be technically challenging and economically infeasible for researchers who lack specific expertise in molecular biology, bioinformatics and/or have limited access to expensive computational resources or sophisticated and often dedicated instrumentation. Thus, the need remains for simple and extensible methods for generating genome-scale variation. An alternative class of methods focuses on amplicon-based sequencing (Campbell et al., 2018; Eriksson et al., 2020). One such method is multiplexed ISSR genotyping by sequencing (MIG-seq; Suyama & Matsuki, 2015), which generates amplicons using primers comprising simple sequence repeat (SSR) motifs in multiplex, and sequences the inter-simple sequence repeat (ISSR) fragment ends. Briefly, ISSR involves the use of primers that match various microsatellite regions in the genome (Zietkiewicz et al., 1994); the primers often include 1–3 bp specific or degenerate anchors to preferentially bind to the ends of these repeat motifs (Gupta et al., 1994; Zietkiewicz et al., 1994) or consist entirely of an SSR motif (Bornet & Branchard, 2001). MIG-seq uses 'tailed,' barcoded, ISSR-motif primers in a two-step PCR protocol, first with the amplification of ISSR, and the second with common primers matching the tails to enrich these amplicons, followed by size selection and sequencing on the Illumina platform. This method has the advantage of not requiring more conventional Illumina library preparation (i.e. fragmentation of genomic DNA and ligation of barcoded sequencing adapters), as the adapters and barcodes are included in the tailed ISSR-motif primers, thus reducing the cost and labour associated with conducting many library preps.

While MIG-seq has been cited by more than 90 studies to date (e.g. Eguchi et al., 2020; Gutiérrez-Ortega et al., 2018; Park et al., 2019; Takata et al., 2019; Tamaki et al., 2017), many questions remain with regard to its efficiency and reproducibility. First, the use of long, tailed, ISSR primers raises the possibility of primer multimerization and unpredictability of binding specificity (eight forward and eight reverse primers in multiplex, as has typically been implemented). Second, as originally implemented, the protocol requires 96 unique forward-indexed primers, each 61 bp in length. The cost of synthesizing such lengthy primers equates to thousands of US dollars spent up-front, which may be prohibitive for many researchers, though these costs could be mitigated somewhat via dual indexing and the use of shorter adapter sequence tails. Third, MIG-seq only produces sequence data from the ISSR amplicon fragment ends, for example as is done with ribosomal DNA metabarcoding, potentially missing significant levels of variation by not sequencing the entire amplified fragments. The numbers of single nucleotide polymorphisms (SNPs) reported in studies using MIG-seq range from a few hundred in intraspecific studies (e.g. Suyama & Matsuki, 2015) to thousands in a species-level phylogenetic study (Eguchi et al., 2020), although missing data comprised the majority (81.4%) of the SNP matrix generated via the latter study. Data of this quantity and completeness can be useful for basic population genetics or phylogeographical studies, but are inadequate for studies requiring densely sampled polymorphisms across the genome (e.g. Quantitative Trait Locus mapping, genomic scans of adaptive variation, pedigree analysis). Indeed, Suyama and Matsuki (2015) state: '...the number of SNPs is fewer in our method [than in RAD-seq] (e.g. ~1,000 vs. ~100,000 SNPs), which means low efficiency in terms of the cost per SNP and sequencing effort'.

Here we present ISSRseq, a novel RRS method that is straightforward, extensible and uses single-primer ISSR amplicons. Briefly, our approach is to produce single-ISSR primer amplicons (as opposed to using tailed, highly multiplexed primer pairs), pool amplicons from multiple primers per sample (Figure 1), conduct low-cost, fragmentase-based Illumina library preparation, and sequence entire ISSR regions on the Illumina platform (Figure 2). Furthermore, we provide a user-friendly set of UNIX BASH scripts that together comprise an analysis pipeline for user-customized data quality control and analysis, output of SNP data in formats commonly used for population genomics and phylogeography, and an easy to use script template for downstream population genomic analyses in R. Our findings demonstrate that ISSRseq can generate comparable levels of variation to RAD or GBS, and orders of magnitude more variation than MIG-seq, while containing low levels of missing data. We present case studies of the utility of this method in *Populus deltoides* W. Bartram ex Marshall, a species with a relatively small and accessible reference genome, and two mycoheterotrophic orchid species with large, uncharacterized genomes: *Corallorhiza bentleyi* Freudenst. and *C. striata* Lindl. All laboratory research was carried out by undergraduate students, demonstrating that this method is amenable and accessible to those with relatively limited experience in molecular biology.

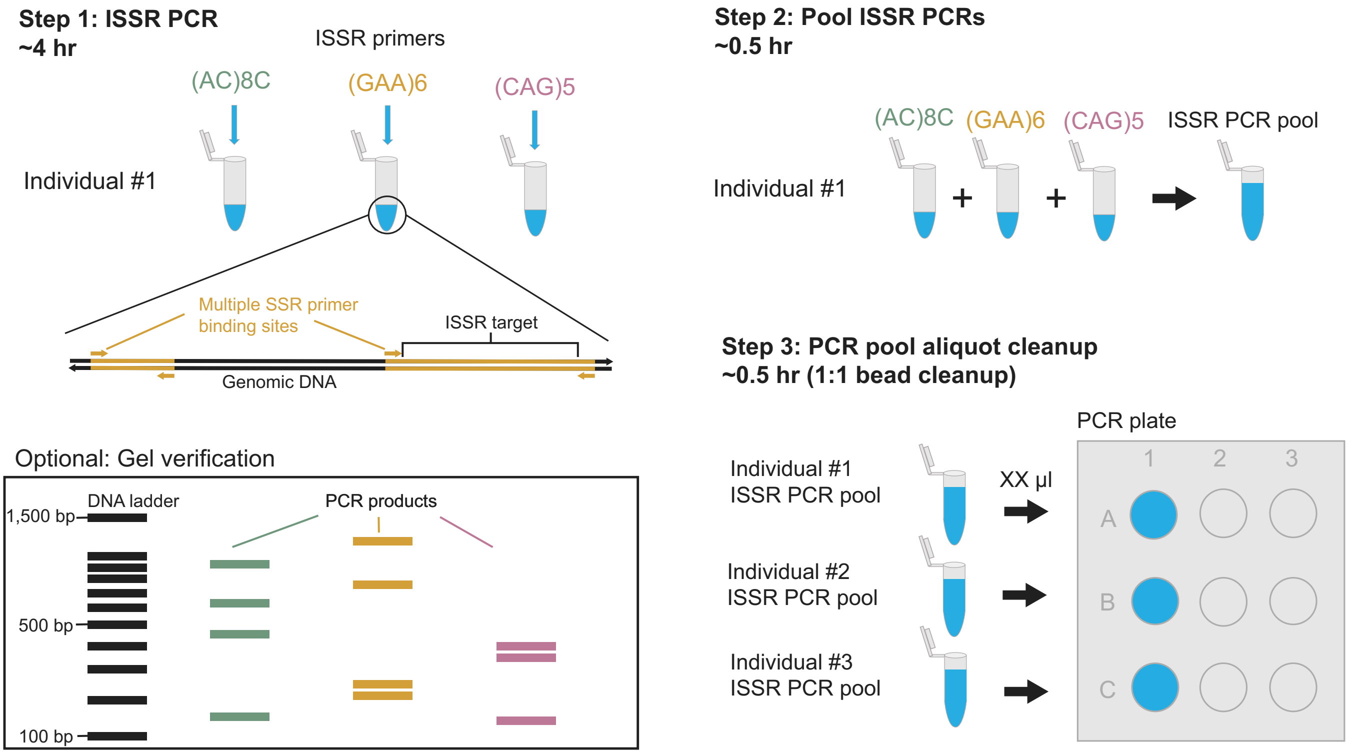


FIGURE 1 Critical steps of ISSRseq polymerase chain reaction (PCR), including the approximate time projected for library preparation of 48 samples with four primer sets. The optional gel verification step depicts three different ISSR primers for a single individual. PCR pool aliquot in Step 3 dependent on number of primer sets chosen to create library

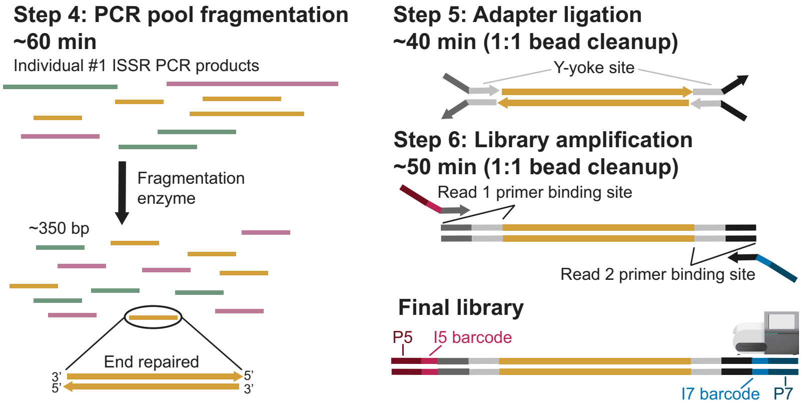

Step 4: PCR pool fragmentation
~60 min
Individual #1 ISSR PCR products
Fragmentation enzyme
~350 bp
End repaired
3'
5'
5'
3'
Step 5: Adapter ligation
~40 min (1:1 bead cleanup)
Y-yoke site
Step 6: Library amplification
~50 min (1:1 bead cleanup)
Read 1 primer binding site
Read 2 primer binding site
Final library
P5
I5 barcode
I7 barcode
P7

FIGURE 2 Critical steps of ISSRseq library preparation, including the approximate time projected for library preparation of 48 samples with four ISSR primers

2 | MATERIALS AND METHODS

2.1 | Laboratory methods

2.1.1 | Sample collection and DNA extraction

Our experimental design was chosen to evaluate the performance of ISSRseq at differing taxonomic scales. We selected *C. bentleyi* to determine whether our method was informative below the level of species, *C. striata* to evaluate the performance across putative species boundaries, and *P. deltoides* to test ISSRseq within a single accession of a species with a well-characterized genome. We collected 87 individuals from 28 *C. striata* localities, as in Fama et al. (2021), Barrett and Freudenstein (2011) and Barrett et al. (2018; Table 1), and 37 individuals from six *C. bentleyi* localities. Samples of *C. striata* were collected from 10 US states and two Canadian provinces (British Columbia and Manitoba, Table 1). Sampling of *C. striata* was focused on the western USA and in particular on California, where the taxonomic status of populations in this complex is uncertain (Barrett et al., 2011, 2018). Samples of *C. bentleyi* were collected in Allegheny, Bath and Giles Counties, Virginia, USA; and from Monroe County, West Virginia, USA (Table 1). For both species of *Corallorhiza*, approximately 0.2 g of perianth or ovary tissue was removed with a sterile scalpel (so as not to include seed material) and DNA was extracted using a CTAB DNA extraction protocol, modified to 1/10 volume (Doyle & Doyle, 1987). A negative control reaction using ultrapure water and no template DNA was included for each set of single-primer reactions. We collected leaf tissue of *P. deltoides*

TABLE 1 Sampling regions and localities for *C. striata* individuals and sampling counties and localities for *C. bentleyi* individuals. United States and Canadian provinces are abbreviated

Region/county	Sampling locality	Number of individuals
C. striata		
Coast Ranges Cascades	Ashland County, OR	2
	Lane County, OR	2
	Marin County, CA	5
	Santa County, Cruz, CA (1)	3
	Santa County, Cruz, CA (2)	4
	Sonoma County, CA (1)	2
	Sonoma County, CA (2)	1
Sierra Nevada	Fresno County, CA	7
	Mariposa County, CA	3
	Nevada County, CA	1
	Placer County, CA	5
var. *striata* N USA	Cache County, UT	5
	Idaho County, ID	4
	Lewis and Clark County, MT	5
	Lewis County, WA	2
	Natrona County, WY	4
	Skamania County, WA	2
	Thompson-Nicola Regional District, BC	2
	Winnipeg Region, MAN	2
var. *vreelandii* SW USA	Graham County, AZ	4
	Otero County, NM	4
	Ouray County, CO	5
	Tooele County, UT	5
	Utah County, UT	4
C. bentleyi		
RG	Alleghany County, VA	5
CG	Bath County, VA	14
BS	Giles County, VA	3
OR	Giles County, VA	5
WR	Giles County, VA	5
PM	Monroe County, WV	5

accession WV94 from a plantation located at the West Virginia University Agronomy Farm (39.658889–79.905278) in Morgantown, West Virginia (Macaya-Sanz et al., 2017). Genomic DNA was extracted using a DNeasy plant mini kit (Qiagen, Cat. No. 69104).

2.1.2 | ISSR primers and PCR conditions

In all, 21 ISSR primers were selected from the UBC Primer Set #9 (University of British Columbia, Canada; paper available on GitHub, www.github.com/btsinn/ISSRseq), with the addition of five unanchored primers designed by the authors (Table 2). We used only eight of these primers in our study of *C. striata* to test extensibility of our method with regard to the number of ISSR primers used to generate amplicons for sequencing. After several rounds of initial optimization, reactions were set up in 10 μl volumes with 5 μl 2x Apex PCR Master Mix (Genesee Scientific; Cat. No. 42-134), 0.5 μl 5 M Betaine (Fisher Scientific; Cat. No. AAJ77507AB), 1.0 μl of each single primer at 10 μM starting concentration (one primer per reaction; Integrated DNA Technologies), 2.5 μl nuclease-free ultrapure water and 1.0 μl template DNA (diluted to 20 ng/μl prior to PCR with Tris-EDTA pH 8.0). PCRs were set up as single master mixes with individual ISSR primers and aliquoted to 96-well PCR plates. Conditions were as follows for each primer: 5 min at 95°C; 30 cycles of 95°C (30 s), 50°C (30 s) and 72°C (1 m); and a final extension of 10 m at 72°C. Caution was taken during PCR amplification to avoid human-, microbial- or plant-based contamination; all laboratory surfaces were disinfected with 10% sodium hypochlorite solution prior

TABLE 2 List of ISSR primers used in this study for the three test species. Numbers listed under 'ISSR primer' correspond to the UBC Primer Set 9 names (paper available on GitHub www.github.com/btsinn/ISSRseq)

ISSR primer	Motif	*P. deltoids*	*C. bentleyi*	*C. striata*
813	(CT)8T	X	X	
814	(CT)8A	X	X	
815	(CT)8G	X	X	
817	(CA)8A	X	X	
820	(GT)8T	X	X	
824	(TC)8G	X	X	
826	(AC)8C	X	X	X
834	(AG)8YT	X	X	X
836	(AG)8YA	X	X	X
840	(GA)8YT	X	X	X
843	(CT)8RA	X	X	
845	(CT)8RG	X	X	
848	(CA)8RG	X	X	
855	(AC)8YT	X	X	
856	(AC)8YA	X	X	X
857	(AC)8YG	X	X	
858	(AC)8RT	X	X	
859	(TG)8RC	X	X	
860	(TG)8RA	X	X	
868	(GAA)6	X	X	X
873	(GACA)4	X	X	
caa5	(CAA)5	X	X	X
cag5	(CAG)5	X	X	X
gtt5	(GTT)5	X	X	
gat5	(GAT)5	X	X	
cat5	(CAT)5	X	X	

to amplification. PCR products were visualized on 1% agarose gels to verify reaction success (see Figure S1 for an example gel image).

2.1.3 | Library preparation and sequencing of ISSR amplicon pools

PCR products were pooled for individual accessions across all primer amplification reactions, and the remaining volume of original reactions was stored in a −80°C freezer as backups. Pooled PCR products were cleaned of excess PCR reagents via Axygen® AxyPrep FragmentSelect-I Kit (Corning, Cat. No. MAG-FRAG-I-50) and two 80% ethanol washes on a magnetic plate. Cleaned PCR pools were quantified via Nanodrop spectrophotometry and diluted with TE buffer to 5 ng/μl. Library preps were conducted with the QuantaBio sparQ DNA Frag and Library Prep Kit (QuantaBio, Cat. No. 95194-096), a relatively inexpensive, rapid library kit that uses a fragmentase to shear genomic or amplified DNA. The library preparation protocol is described in detail elsewhere (www.github.com/btsinn/ISSRseq). Briefly, we scaled library preparation volumes by half (thus a 96-reaction kit yields 192 preps). The sparQ Frag and Library Prep Kit allows fragmentation and end repair in a single step followed by Y-yoke adapter ligation (Glenn et al., 2019). We then performed a magnetic bead cleanup, total library amplification with barcoded Illumina iTru primers (Glenn et al., 2019) and a final bead cleanup; the bead to sample ratio was 1:1 for both cleanups. Fragmentase time was optimized in earlier trials, and consistently gave the best target library sizes for Illumina sequencing at 3.5 min (by contrast, for high molecular weight, genomic DNA fragmentation is suggested to be set at 14 min). Three microliters of each library were run on a 1% agarose gel with GeneRuler 1 kb Plus DNA Ladder (ThermoFisher Scientific; Cat. No. SM1331) followed by quantification via Qubit™ dsDNA BR Assay Kit (Invitrogen; Cat. No. Q32850). Amplicons for each accession were pooled at equimolar ratios. Size selection of an aliquot of the final pool was conducted using magnetic beads at a bead to sample ratio of 1:1, but users could also conduct two selective bead cleanups with different sample:bead ratios or gel excision. Fragment size range and intensity were quantified on an Agilent 2100 Bioanalyzer (Agilent Technologies) and final library quantification was conducted via quantitative PCR at the West Virginia University Genomics Core Facility. The final library size ranged from 250 to 550 bp. Pooled, indexed ISSR amplicons were sequenced using 2 × 150 bp Illumina MiSeq (reagent kit v2) for *C. bentleyi* and *P. deltoides*. For *C. striata*, three sequencing runs were conducted: one lane each of 2 × 100 bp and 2 × 50 bp on a HiSeq1500 at the WVU-Marshall Shared Sequencing Facility, and one run of 2 × 150 bp (reagent kit v2) Illumina MiSeq at the WVU Genomics Core Facility.

2.2 | Bioinformatics and downstream analyses

Our bioinformatic analysis pipeline consists of five BASH scripts for use on UNIX-based systems, each with customizable options, allowing simple but flexible parameter adjustment to fit the needs of a particular dataset or project (Figure S2). We also supply an R script template to conduct various population genomic analyses. These scripts are not meant to dictate how ISSRseq data are treated or analysed by users, but rather they are meant to serve as an accessible example of analysis potential for these data. Below, we detail the workflow iteratively and in the order of script usage. One of the BASH scripts, ISSRseq_ReferenceBased.sh, is optional if the user wishes to map to a reference genome, previously assembled contigs, or to a previously assembled set of ISSRseq amplicons. BASH and R scripts are provided and a detailed wiki with usage examples can be found via the ISSRseq GitHub repository wiki (www.github.com/btsinn/ISSRseq/wiki).

(1) ISSRseq_AssembleReference.sh

Read pre-processing

BBDUK (38.51, Bushnell, 2020) is used to remove adapters and priming sequences from reads, quality trim, and exclude GC-rich or -poor reads for each sample. Additionally, we used the kmer trimming feature of BBDUK to remove ISSR motifs used as primer sequences from the ends of reads. Kmer length was set to 18 and the 'mink' flag set to 8. The use of the 'mink' flag allowed for the removal of matched kmers down to a length of 8 bp from that of the supplied priming sequences (Table 2). We also enabled trim by overlap ('tbo') and 'tpe,' which in tandem allowed for adapter trimming by leveraging read overlap and in such an event trimmed both read pairs to the same length to ensure adapter removal. Trimmed reads for which the average quality score was below 10 or length was less than 50 bp were excluded, with the exception of the HiSeq 50 bp reads. Hard trimming of read ends was not conducted when combining MiSeq 100 and 150 bp, and HiSeq 50 bp data generated for *C. striata*.

Reference assembly

We then used ABySS-pe (version 2.2.4, Jackman et al., 2017) to assemble the trimmed reads from the user-specified reference sample using a kmer length of 91. Kmer choice was guided by our desire to minimize potential assembly errors due to the presence of low-complexity repeats (SSR motifs). BBDUK was then used to trim the assembled contigs in the same fashion as was used for read trimming, but with the GC content filter set to 35% and 65% and with the entropy filter enabled and set at 0.85. A reference index and sequence dictionary were created by SAMtools 'faidx' (version 1.7-13-g8dee2a2, Li et al., 2009) and the Picard tool 'CreateSequenceDictionary' (version 2.22.8; Broad Institute), respectively.

Contaminant filtering

Next, we used the trimmed and filtered reference contigs as queries in BLASTn (version 2.6.0, Camacho et al., 2009) to identify putative contaminant loci by using the human genome, and those of 153 genomes of organisms that could be expected as common contaminants of plant samples (see Supporting Information), as subjects with an e-value cut-off of 0.00001. Contigs with e-values below this cutoff are excluded from the final reference. The user also specifies a plastid genome, and/or any other genome of interest, to be used as a negative reference. Contigs that can be mapped to the negative reference by BBMap (version 38.51, Bushnell, 2020) were also excluded from the reference assembly. Contigs identified as putative contaminants or representing the negative reference are written to a FASTA file.

(2) ISSRseq_CreateBams.sh

Read mapping and BAM creation

BBMap (version 38.51) was also used to map trimmed reads from each sample to the assembled contigs using default mapping settings and killbadpairs enabled. SAMtools was used to sort and index the BAM (Cock et al., 2015) files of each sample. We then used the Picard tools, MarkDuplicates (version 2.22.8; Broad Institute), to mark PCR and optical duplicate reads, and BuildBamIndex (version 2.22.8; Broad Institute) to re-index these final BAM files.

(3) ISSRseq_AnalyzeBAMs.sh

Variant calling and filtering

We used the state-of-the-art variant calling pipeline GATK4 (version 4.1.8, McKenna et al., 2010) to call, filter and jointly score variants among all samples simultaneously. HaplotypeCaller (Poplin et al., 2017) identified potential variant sites and called variants from locally reassembled portions of each BAM, with --linked-de-bruijn-graph, --native-pair-hmm-use-double-precision and -ERC GVCF modes enabled. GVCF files from each sample were then combined into a single VCF file with CombineGVCFs. GenotypeGVCFs was then used to perform joint variant scoring on the combined VCF file. Scored variants for downstream analysis were restricted to biallelic SNPs and INDELs, which were then hard-filtered guided by GATK Best Practices hard filtering recommendations (DePristo et al., 2011): 'AF >0.01 && AF < 0.99 && QD > 2.0 && MQ > 40.0 && FS < 60.0 && SOR < 3.0 && ReadPosRankSum > −8.0 && MQRankSum > −12.5 && QUAL > 30.0'. Since the workflow exists as a BASH script, users can customize any of the variant filtering parameters to suit their needs.

(4) ISSRseq_CreateMatrices.sh

Matrix creation

VCF2PHYLIP (version 2.0; Ortiz, 2019) was used to coerce the minor allele and hard-filtered SNP variants into nexus, phylip and binary SNP formats with varying matrix inclusion thresholds corresponding to the minimum number of samples in which a variant was identified.

(5) ISSRseq_ReferenceBased.sh [Optional]

This script processes input reads and prepares the necessary file structure for the use of the pipeline with a pre-existing reference, for example, if the user has a sequenced genome or previously generated de novo assembly of contaminant-filtered ISSR amplicons at their disposal. Unlike ISSRseq_AssembleReference.sh, this script does not conduct contaminant filtering or trim both ends of the sequence reads since the input reads are not used for de novo assembly of the reference. Users should remove organellar and other non-target contigs from the reference prior to using this script. The output directory can then be used for steps 2–4 outlined above.

2.3 | Reference, coverage and missing data comparisons for *C. striata* datasets

To test the effects of using different Illumina sequencing strategies (i.e. sequencing effort and read length), we collected three datasets for the same 87 accessions of *C. striata*: Illumina MiSeq 2 × 150 bp, HiSeq 2 × 50 bp and HiSeq 2 × 100 bp (see above). We were specifically interested in the total number of SNPs, mean coverage depth and percent missing data for each of the three individual datasets plus a 'combined' dataset using all three. We mapped reads as above to a common reference based on the combined dataset, with the goal of producing the most complete reference for mapping reads from individual datasets.

2.4 | Population genetic analyses in R

All population genetic analyses were carried out using packages in the R software environment (R Core Team, 2019). We describe each of the analyses and the prerequisite data preparation and population strata used in the subsections below. An example R script is provided in the ISSRseq GitHub repository (www.github.com/btsinn/ISSRseq) and a walkthrough of these analyses is provided via the wiki (https://github.com/btsinn/ISSRseq/wiki/ISSRseq-R-Analyses).

2.4.1 | Data preparation

Prior to population genetic analyses, we thinned our hard-filtered VCF files to retain one variant per locus by setting the *thin* flag of VCFtools (0.1.15; Danecek et al., 2011) to the maximum contig length, and then used the *max-missing* flag to remove variants for which missing data exceeded 90% and 80% for *C. bentleyi* and *C. striata*, respectively.

2.4.2 | Analysis of molecular variance (AMOVA) and *F*-statistic estimation

We used AMOVA to assess population structure by partitioning the genetic variation based on predefined population categories using the POPPR package (Kamvar et al., 2015; '*poppr.amova*' function). We performed a permutation test for 1,000 iterations to test phi statistics for significance ('*randtest*' function). We also estimated commonly used population statistics including observed heterozygosity (H_o), observed gene diversities (H_s), inbreeding coefficient (F_{is}) and fixation index (F_{st}) across all loci using the HIERFSTAT package (Goudet, 2005; '*basic.stats*' function) for both species. The HIERFSTAT package was also used to estimate pairwise F_{st} ('*genet.dist*' function) among population categories.

2.4.3 | Principal components analysis

We used the ADEGENET package (Jombart & Ahmed, 2011) to perform a principal components analysis (PCA) and discriminant analysis of principal components (DAPC) to assign subpopulation membership probabilities to individual plants ('*dapc*' function).

3 | RESULTS

3.1 | Sequencing and reference assemblies

3.1.1 | *C. striata* complex

PCR using eight SSR primers (Table 2) successfully generated amplicons from 81 of the 87 accessions. The total combined sequencing read pool comprised 310,379,211 read pairs, of which 239,142,938 remained after trimming and filtering (Table 3). We selected accession 253e_CA as the reference individual since it is the sample for which we recovered the greatest number of reads (28,935,452 read pairs). De novo assembly resulted in 17,143 contigs longer than 100 bp, comprising a reference assembly totalling 3,628,930 bp with an N50 of 212 bp and GC content of 47.24%. Putative contaminant filtering excluded 555 contigs that either mapped to a genome of a putative contaminant or to any plastome sequenced from the *C. striata* complex (JX087681.1, NC_040981.1, MG874039.1, NC_040978.1) or that of *C. bentleyi* (NC_040979.1). The reference assembly comprised 14,164 contigs shorter than 250 bp, 450 contigs longer than 500 bp, and 34 contigs longer than 1 kb; the longest contig was 2,032 bp. Contigs of putative contaminant or plastid loci

TABLE 3 Comparison of the analysis of Illumina MiSeq 150 bp PE, and Illumina HiSeq 50 and 100 bp PE reads analysed singly and in combination using the same de novo reference assembly for the *C. striata* complex. An asterisk denotes statistics which were calculated after the removal of six samples which produced few sequences. SNP = single nucleotide polymorphism; bp/SNP = Total base pairs of reference sequence per SNP; *SD* = standard deviation; # min = the minimum number of samples within which a SNP was scored to be included in the data matrix. Variant implies both INDELS and biallelic SNPs

	MiSeq 2 × 150	HiSeq 2 × 50	HiSeq 2 × 100	Combined
# raw read pairs/# post-trim	9,899,895/13,435,422	151,025,536/97,981,594	149,453,780/138,080,666	310,379,211/239,142,938
Total HaplotypeCaller variants	249,697	328,726	560,669	641,667
Filtered SNPs	8,177	12,694	28,401	25,904
Mean filtered variant QUAL score	1,636.36	499.96	2,982.01	4,685.16
bp/SNP (total filtered SNPs)	410.42	280.03	122.10	140.10
Mean coverage depth/ locus/accession	3.81	5.55	10.67	14.82
SD coverage depth	3.48	7.30	12.42	16.71
Mean coverage depth/ locus/accession*	4.09	5.94	11.44	15.88
SD coverage depth*	3.44	7.41	12.52	16.82
SNPs/%missing data (min 1)*	8,177/54.8%	12,694/51.1%	28,401/42.7%	25,904/40.0%
SNPs/%missing data (min 10)*	7,219/49.5%	11,808/47.9%	26,506/38.9%	24,168/35.9%
SNPs/%missing data (min 50)*	2,412/20.7%	3,991/19.7%	13,387/18.2%	13,762/17.7%
SNPs/%missing data (min 80)*	261/3.1%	627/2.3%	3,209/2.1%	3,503/2.1%

totalled 144,068 bp with an N50 of 230 bp, GC content of 44.17%, and a maximum contig length of 1,399 bp. This reference assembly was used for all comparisons of variant scoring for *C. striata*, below. For sample 253e_CA, the average insert size was 290 bp, 65.6% of reads mapped to the final reference assembly and 24.1% of reads mapped to the putative contaminant assembly. The mean coverage depth of filtered variants scored in accession 253e_CA using this reference was 123.29x.

3.1.2 | *C. bentleyi*

Amplicons were successfully generated from 37 *C. bentleyi* accessions using 26 primers shown in Table 2. The total combined sequencing read pool comprised 40,949,418 read pairs, of which 25,523,858 survived trimming and filtering. Accession B8 was chosen as the reference individual, as it was the accession for which we recovered the greatest number of reads (913,462 read pairs). De novo assembly recovered 16,813 contigs longer than 100 bp, comprising a reference assembly totalling 3,928,602 bp in length with an N50 of 234 bp and GC content of 46.43%, excluding 686 contigs that either mapped to a genome of a putative contaminant or to the plastome of *C. bentleyi* (NC_040979.1). The majority of reference contigs were less than 250 bp in length (11,894), while 3,748 were longer than 250 bp, 1,003 were longer than 500 bp and 168 were longer than 1 kb; the longest contig was 2,390 bp. Putative plastid or contaminant contigs totalled 259,774 bp with an N50 of 463 bp, GC content of 43.24% and a maximum contig length of 2,436 bp. For sample B8, the average insert size was 421 bp, with 37.5% of reads mapped to the final reference assembly and 13.3% mapped to the putative contaminant assembly. The mean coverage depth of filtered variants scored in accession B8 using this de novo reference was 16.42x.

3.1.3 | *Populus deltoides* WV94

We used all 26 SSR primers used in *C. bentleyi* to conduct PCR amplification of one *P. deltoides* clone (WV94), which was multiplexed along with *C. bentleyi* accessions. The raw sequencing pool comprised 589,403 read pairs, of which 390,992 survived trimming. Kmer trimming of adapter and/or SSR primer sequences occurred on 51.84% of reads. The chromosome-level assembly of *P. deltoides* (445, version 2.0; https://phytozome.jgi.doe.gov) comprises 403,296,128 bp. Trimmed reads covered 4,376,825 bp or 1.09% of the genome (55.96% of trimmed reads), covering a median of 1.04% (range = 0.62% to 1.85%; *SD* = 0.34) of each chromosome (Table S1). Visual observation of BAM files qualitatively suggested that SSRs were amplified relatively evenly throughout the genome (Figure S3). A positive correlation between chromosome length and percent of each chromosome covered by mapped trimmed reads was recovered by linear regression ($r^2 = 0.412$, p-value = 0.003; Figure S4), congruent with our visual assessment of SSR amplification. Contrary to this positive correlation was read mapping to Chromosome 8, where 53,666 reads covered 1.85% of its total length, the largest of such values, despite its standing as the eighth longest chromosome. Mean coverage depth of filtered variants across all chromosomes was 36.71x.

3.2 | Read mapping and variant calling results

3.2.1 | *C. striata*

Analysis of individual MiSeq and HiSeq sequencing runs of PCR amplicon pools and their combination resulted in variable numbers of raw variants, variant quality scores, variant coverage depth, and number and missingness of final filtered variants (Table 3). Exclusion of six samples, for which sequencing produced few reads, increased filtered variant depth for the MiSeq 150 run to 4.09 and that of the combined read pool to 15.88. Higher variant quality, coverage and concomitant reduction in missingness of data guided our decision to use the analysis of the combined read pool for downstream analyses. In all, 24,078 SNPs remained after VCFtools missingness filtering, and thinning to 1 variant per locus left 6,589 variants in the final *C. striata* matrix for analysis in R.

3.2.2 | *C. bentleyi*

HaplotypeCaller identified 236,694 total variants. SNPs comprised 47,851 of filtered variants or 76.11 bp/SNP. The mean coverage and quality score of filtered variants across all samples were 9.75 and 314.13, respectively. Although the mean SNP quality score among the *C. bentleyi* samples was lower than that of SNPs scored from *C. striata*, missing data comprised only 10% of total filtered SNPs, 8.6% scored from at least 10 accessions, 7.0% scored from a minimum of 20 accessions and 4.5% scored from 30 of the 37 accessions. After filtering with VCFtools, 51,132 variants remained in the *C. bentleyi* matrix and 3,536 variants remained after thinning for analysis in R.

3.2.3 | *P. deltoides* WV94

Of the 8,134 variants called by HaplotypeCaller in the *P. deltoides* WV94 clone we sequenced, 1,040 SNPs passed hard filtering. Variants were identified on all chromosomes (Figure S5), and all filtered variants were heterozygous.

3.3 | AMOVA and population statistics

For the *C. bentleyi* dataset, which contained 3,536 variants, we found that 92.2% of genetic variation was explained within individuals which was significantly less than expected by chance (phi = 0.079, sigma = 23.1, p-value < 0.001), while 4.45% was explained among sampling localities within county which was significantly greater than

TABLE 4 Analysis of molecular variance (AMOVA) output with fixation index for *Corallorhiza* populations. # loci = number of variants analysed; columns 3–5 are percent variation explained by each hierarchical level; 'a/m regions-county' = among regions (*C. striata*) or county (*C. bentleyi*); 'a/m loc' = among sampling localities within region or county; and 'w/in ind' = within individuals. Columns 6–8 are values of fixation index (Φ) and their significance: *$p < 0.005$, **$p < 0.001$. 'ΦCT' = among regions or county; 'ΦSC' = among sampling localities within regions or county and 'ΦIT' = within individuals

Species	# Variants	b/w region-county	b/w loc	w/in ind	Φ_{CT}	Φ_{SC}	Φ_{IT}
C. bentleyi	3,536	3.44	4.45	92.2	0.034	0.046**	0.079*
C. striata	6,589	49.2	8.24	42.5	0.492**	0.162**	0.575**

expected by chance (phi = 0.046, sigma = 1.11, *p*-value < 0.002), and 3.44% of variation was explained among county (phi = 0.034, sigma = 0.860, *p*-value = 0.314; Table 4). Average observed heterozygosity and observed gene diversity were 0.066 and 0.078, respectively (Table 5). The *F*-statistic coefficients for the *C. bentleyi* sampling locality grouped populations that had low inbreeding coefficients and genetic distances (Figure 3a).

For *C. striata*, we found that 42.5% of genetic variation was explained within individuals which was significantly less than expected by chance (phi = 0.575, sigma = 28.2, *p*-value < 0.001), while 8.24% was explained among sampling localities (phi = 0.162, sigma = 5.46, *p*-value < 0.001), and 49.2% of variation was explained among regions (phi = 0.492, sigma = 32.6, *p*-value < 0.001) which were both significantly greater than expected by chance (Table 4). Average observed heterozygosity and observed gene diversity were 0.077 and 0.096, respectively. The *F*-statistic coefficient values were moderately high for the *C. striata* when treating geographical region at the level of subpopulation (Table 5; Figure 3b).

3.3.1 | PCA results

We retained three PC axes for the *C. bentleyi* analysis which explained a total of 14.1% of the variation in the genetic dataset (Figure 4a). Discriminant analysis of principal components revealed the number of clusters appropriate for assigning membership probability based on principal components was two for *C. bentleyi* (Figure 4b). In the *C. striata* analysis, we retained four PC axes which explained 32.0% of the total genetic variation (Figure 4c). *C. striata* sample membership was best explained by four clusters (Figure 4d).

4 | DISCUSSION

4.1 | *C. striata*

Population genetic analyses of ISSRseq-generated SNPs correspond well to previously published phylogeographical patterns recovered using targeted sequence capture of plastid loci (Barrett et al., 2018) and population genetic estimates using nuclear markers (Barrett & Freudenstein, 2011). For example, on the basis of three nuclear introns, Barrett and Freudenstein (2011) estimated mean Φ_{CT} as 0.450, using some of same accessions used in the present study, and we estimated the same parameter at 0.492 using 6,589 variants generated by ISSRseq. Furthermore, F-coefficients and the results of AMOVA and DAPC suggest the presence of population subdivision and geographical partitioning of genetic variation in like fashion with that found using previous AMOVA, STRUCTURE analyses (Barrett & Freudenstein, 2011) and phylogenomic inference (Barrett et al., 2018). The congruence of the results of independent analyses using thousands of SNPs identified using ISSRseq with those of previous studies suggests that these data found using our novel method (a) are appropriate for commonly used analyses; (b) contain reliable population genomic signal and (c) are preferable since ISSRseq generates genome-scale data without prior knowledge of the genome.

TABLE 5 *Corallorhiza* population statistics table for *C. bentleyi* sampling locality and *C. striata* region. # loci = number of variants analysed, H_o = observed heterozygosity, H_s = observed gene diversities, F_{is} = inbreeding coefficient, F_{st} = fixation index

Species	# Loci	H_o	H_s	F_{is}	F_{st}
C. bentleyi	3,536	0.066	0.078	0.149	0.000
C. striata	6,589	0.077	0.096	0.199	0.198

4.2 | *C. bentleyi*

ISSRseq corroborated results of a traditional ISSR investigation study of *C. bentleyi* conducted previously by our group. Fama et al. (2021) scored ISSR bands visually for two primers used in this study and found that 89% of molecular variance occurred within populations. Using ISSRseq, we likewise found that the majority of genetic variation was explained within sampling localities, with only 4.45% of the genetic variation explained by sampling locality. Additional evidence of the recovery of population genomic signal in these data was our identification of fixed homozygous sites exclusive to cleistogamous individuals. We find the general congruence of visually scored ISSR banding patterns in Fama et al. (2021) with the more sensitive, sequence-based nature of ISSRseq to be additional corroboration of our new method.

4.3 | *Populus deltoides* WV94

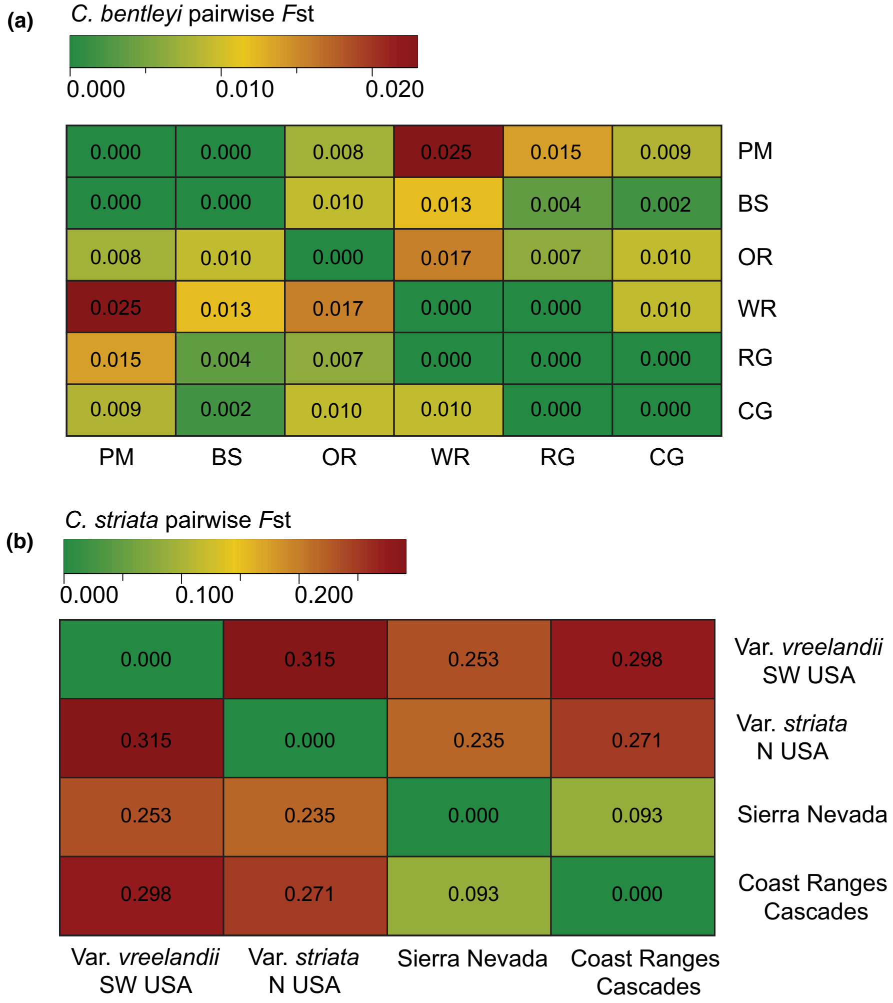
Our analysis of a *Populus deltoides* WV94 clone demonstrated that the loci sequenced and variants identified by ISSRseq are located throughout the genome, and variants were not obviously clustered in particular chromosomes or their centromeres or telomeres (Figures S3 and S5). As expected, we observed that the depth of final filtered SNPs was greater (36.71x) than the median depth of coverage for each chromosome (1.04x), which is to be expected with any reduced representation sequencing method.

4.4 | ISSRseq is straightforward and extensible

ISSRseq generates genomic variants on a scale that is comparable to other established RRS methods while minimizing time and wet-laboratory complexity. Assuming the availability of two PCR machines and familiarity with basic wet-laboratory techniques, users can go from extracted DNA to an Illumina sequencing library for 48 samples and 4 primer sets within an 8-hr workday. This timeframe and capacity can be greatly increased for users who use 96- or 384-well plates to conduct PCR.

ISSRseq is suitable for users with minimal laboratory experience or resources, since only a thermocycler and commonplace equipment such as a DNA quantitation device and affordable neodymium magnets used for PCR cleanup with magnetic beads are necessary prior to sequencing. Indeed, the ISSRseq data analysed here were generated by undergraduates, and ISSRseq projects have been conducted in an undergraduate course at West Virginia University. Additionally, the use of a commercially available, fragmentase-based sequencing library preparation kit means that users can receive support directly from a company, rather than relying on personal communications with authors. While alternative library preparation protocols can be used with ISSRseq, we found the efficiency and reliability of a commercially available kit that can prepare a library in about 2.5 hr to be fitting for our purposes. A sizable fraction of the cost of ISSRseq, as currently implemented, is the use of a library preparation kit. Our publication of this method stands as a proof of concept, rather than a prescription, and we expect that advanced users of ISSRseq will modify the protocol and analysis pipeline described herein.

The recovery of loci generated among samples sequenced using RAD-like RRS approaches is impacted by mutations to restriction sites, DNA sample impurities and degradation, and user error or random variation during size selection (Andrews et al., 2016). ISSRseq generates loci for downstream analysis via PCR rather than careful size selection of restriction-fragmented DNA via gel excision or pulsed-field electrophoresis, and we expect that locus dropout due to user error or DNA quality will prove to be minimal relative to RRS methods such as ddRAD. In line



	PM	BS	OR	WR	RG	CG
PM	0.000	0.000	0.008	0.025	0.015	0.009
BS	0.000	0.000	0.010	0.013	0.004	0.002
OR	0.008	0.010	0.000	0.017	0.007	0.010
WR	0.025	0.013	0.017	0.000	0.000	0.010
RG	0.015	0.004	0.007	0.000	0.000	0.000
CG	0.009	0.002	0.010	0.010	0.000	0.000

	Var. *vreelandii* SW USA	Var. *striata* N USA	Sierra Nevada	Coast Ranges Cascades
Var. *vreelandii* SW USA	0.000	0.315	0.253	0.298
Var. *striata* N USA	0.315	0.000	0.235	0.271
Sierra Nevada	0.253	0.235	0.000	0.093
Coast Ranges Cascades	0.298	0.271	0.093	0.000

FIGURE 3 Pairwise F_{st} values comparing (a) *Corallorhiza bentleyi* sampling localities and (b) *C. striata* regions. Red indicates high relative differentiation

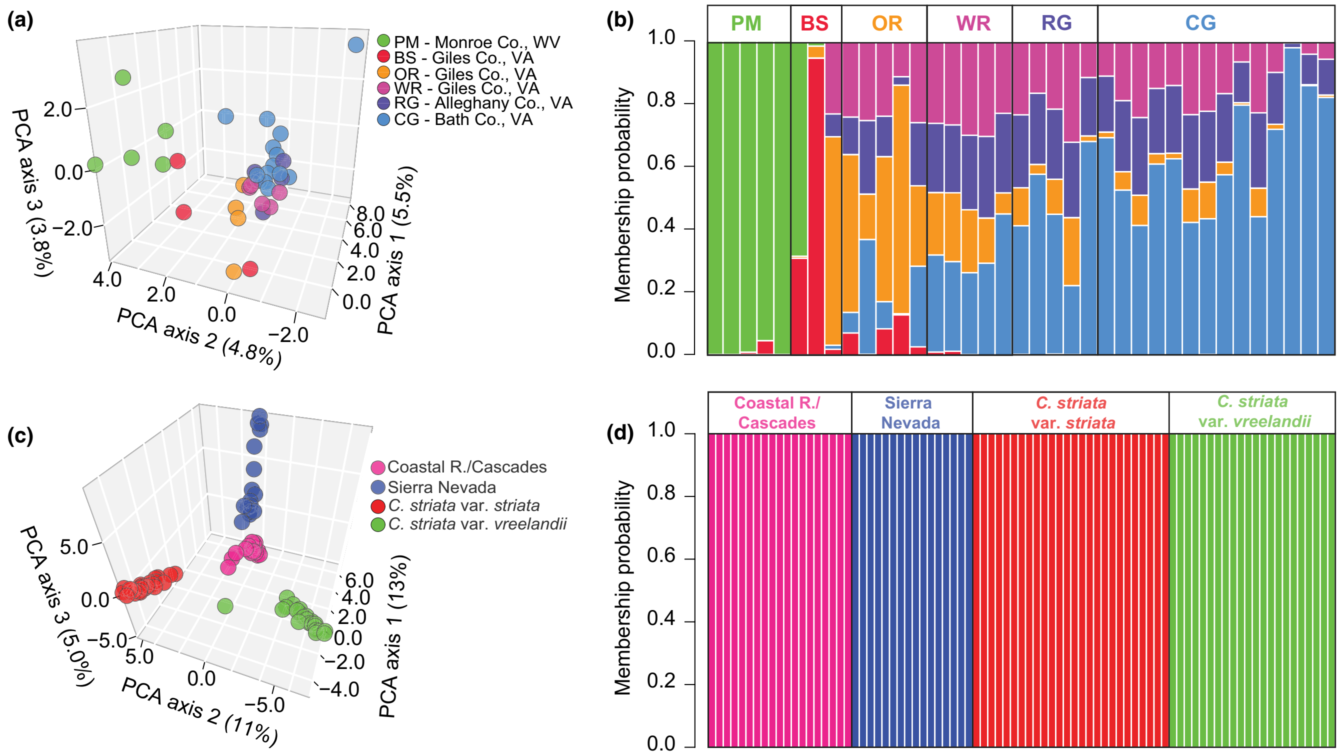


FIGURE 4 Population structure analyses of *Corallorhiza bentleyi* ($n = 37$ accessions) and the *C. striata* complex ($n = 81$ accessions). (a) adegenet principal components analysis (PCA) dotplot and (b) DAPC membership probability barplot for *C. bentleyi* inferred from 51,132 variants. (c) PCA dotplot and (d) DAPC membership probability barplot for the *C. striata* complex inferred from 24,078 variants

with this expectation, missing data accounted for only 10% of the 47,851 SNP data matrix we generated for *C. bentleyi* and only 4.5% missing data when we required a SNP to be called in 30 of 37 individuals. As a point of reference, Tripp et al. (2017) used RADseq to identify 302,987 SNPs in *Petalidium* spp. (Acanthaceae); however, the number of SNPs included in their data matrices (1,568–53,792, depending on parameter choice) was only greater than that recovered using ISSRseq when the missingness cut-off was greater than 60%. While variable SNP recovery may be relatively less of a concern for phylogenetic applications (Eaton et al., 2017; Piwczyński et al., 2020; Tripp et al., 2017), they are known to negatively impact estimation of population genetic parameters. For example, missing data are known to bias estimation of effective population size (N_e) and the inbreeding coefficient (F_{is}; Marandel et al., 2020). The use of PCR for the generation of loci analysed using ISSRseq may mean that locus dropout will be minimized across DNA samples of relatively lower concentration and/or molecular weight than is preferable when using many existing RRS methods.

4.5 | Methodological considerations, suggestions and caveats

Not surprisingly, sequencing depth is an important consideration when using ISSRseq just as with using other RRS methods. Conducting fewer ISSR PCRs using diverse SSR motifs is one option to maximize the efficiency of sequencing depth and number of samples that can be multiplexed during sequencing. For example, we recovered similar cumulative assembled amplicon length in *C. striata* as *C. bentleyi* despite using eight PCRs and 26 PCRs for each, respectively. This is likely due to the fact that many of the primers used for *C. bentleyi* comprised similar SSR motifs but had different anchors, whereas we used fewer primers of differing SSR motifs in *C. striata*, which together reflect the diversity of SSR motifs amplified from the latter species. We also found that 100 bp, paired-end sequencing of the same library on a single HiSeq 2500 nearly tripled the mean sequencing depth per locus (4.09x vs. 11.44x) and more than quadrupled the number of variants scored (8,177 vs. 28,401) over a single lane of paired-end 150 bp sequencing on the MiSeq. Given these results, we suggest that users select primers representing diverse and dissimilar SSR motifs that produce the most amplicons and prioritize read number over read length. Furthermore, we recommend bead-based exclusion of the shortest PCR amplicons prior to library preparation to reduce sequencing of short amplicons likely comprising mostly low complexity sequence. Following these recommendations should also save on PCR reagent cost and worker time.

Although joint variant calling should be able to minimize false positives, we do recommend that future researchers consider conducting more stringent variant filtering that incorporates a control sequence to empirically determine variant filtering parameters. For example, GATK is able to leverage machine learning of known variants to recalibrate variant quality scores and determine appropriate filtering parameters for genomes that are already well characterized (DePristo et al., 2011). The implementation of more sophisticated variant filtration techniques will no doubt reduce noise in ISSRseq data.

The impact of phylogenetic distance or reference choice on locus recovery and variant calling was not tested in the current study. For example, phylogenetic distance between samples and a given reference genome has been shown to differentially impact variant scoring depending on the chosen genotype caller (Duche and Salamin, 2020), a finding that certainly warrants additional study. In future applications of ISSRseq, users are encouraged to explore alternative SNP calling software or approaches for generating data matrices to be used in phylogenomic inference. Regardless, researchers are encouraged to use the BASH scripts provided as a template to guide them in designing a workflow that fits the needs of their particular study.

4.6 | Potential applications

In addition to studies of local adaption, population genomics and phylogenetic inference, ISSRseq could prove a useful technique to those interested in RRS using longer loci, agricultural crop or livestock authentication, forensics and microbiome studies. For example, the use of PCR to amplify loci for sequencing means that the theoretical maximum of locus length is limited only by the DNA polymerase used during the PCR step. By modifying the PCR conditions and reagents, users could conduct long-range amplifications to recover loci that are thousands of base pairs in length. These long loci could facilitate robust estimates of linkage disequilibrium and allele phasing in species without pre-existing genomic resources, and for the construction of gene trees for phylogenomic studies, allowing the use of many multilocus coalescent analyses (ASTRAL, Bayesian species delimitation, etc.). Furthermore, due to the ubiquity of SSRs in genomes, ISSRseq could be used to conduct RRS for the investigation of microbiomes, symbioses and even simultaneous sequencing of hosts and one or more pathogens.

ACKNOWLEDGEMENTS

This work was funded by the WVU Department of Biology, the WVU Eberly College of Arts and Sciences, a WVU Program to Stimulate Competitive Research Grant to C.F.B., NSF Award 1920858 to C.F.B. N.M.F. and M.V.S. were supported by the West Virginia Research Challenge Fund through a grant from the Division of Science and Research, HEPC and in part by (i) the WVU Provost's Office and (ii) the Eberly College of Arts and Sciences. S.J.S. was supported by NSF award 1542509 to S.P.D. We thank the USDA Forest Service for permission to collect samples. We thank Ryan Percifield for technical assistance, and the WVU Genomics Core Facility for sequencing support and services, and CTSI Grant #U54 GM104942 which provides financial support to the WVU Core Facility.

CONFLICT OF INTEREST

The authors are not aware of any conflict of interest.

AUTHORS' CONTRIBUTIONS

C.F.B. conceived the ISSRseq; B.T.S., S.J.S., C.F.B. and S.P.D. designed the laboratory and bioinformatic methodology and protocols; B.T.S., S.J.S., M.V.S. and N.M.F. conducted the laboratory work; C.F.B. conducted all the field work; B.T.S., S.J.S., C.F.B. and S.P.D. analysed the data; B.T.S., S.J.S. and C.F.B. led the writing of the manuscript. All authors contributed to manuscript drafts and provided their consent for publication.

PEER REVIEW

The peer review history for this article is available at https://publons.com/publon/10.1111/2041-210X.13784.

DATA AVAILABILITY STATEMENT

Raw sequencing reads for all samples are available via NCBI BioProject PRJNA771539. Data matrices, scripts used and their usage instructions in wiki format, as well as the wet-laboratory protocol, are available via GitHub (www.github.com/btsinn/ISSRseq) or Sinn et al. (2021).

ORCID

Brandon T. Sinn https://orcid.org/0000-0002-5596-6895
Stephen P. DiFazio https://orcid.org/0000-0003-4077-1590
Craig F. Barrett https://orcid.org/0000-0001-8870-3672

REFERENCES

Altshuler, D., Pollara, V. J., Cowles, C. R., Van Etten, W. J., Baldwin, J., Linton, L., & Lander, E. S. (2000). An SNP map of the human genome generated by reduced representation shotgun sequencing. *Nature*, *407*(6803), 513–516.

Andrews, K. R., Good, J. M., Miller, M. R., Luikart, G., & Hohenlohe, P. A. (2016). Harnessing the power of RADseq for ecological and evolutionary genomics. *Nature Reviews Genetics*, *17*(2), 81–92.

Baird, N. A., Etter, P. D., Atwood, T. S., Currey, M. C., Shiver, A. L., Lewis, Z. A., Selker, E. U., Cresko, W. A., & Johnson, E. A. (2008). Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS ONE*, *3*(10), e3376. https://doi.org/10.1371/journal.pone.0003376

Barrett, C. F., & Freudenstein, J. V. (2011). An integrative approach to delimiting species in a rare but widespread mycoheterotrophic orchid. *Molecular Ecology*, *20*(13), 2771–2786.

Barrett, C. F., Wicke, S., & Sass, C. (2018). Dense infraspecific sampling reveals rapid and independent trajectories of plastome degradation in a heterotrophic orchid complex. *New Phytologist*, *218*(3), 1192–1204.

Bornet, B., & Branchard, M. (2001). Nonanchored inter simple sequence repeat (ISSR) markers: Reproducible and specific tools for genome fingerprinting. *Plant Molecular Biology Reporter*, *19*, 209–215.

Bushnell, B. (2020). bbtools. Retrieved from https://sourceforge.net/projects/bbmap/files/

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: Architecture and applications. *BMC Bioinformatics*, *10*, 421.

Campbell, E. O., Brunet, B. M. T., Dupuis, J. R., & Sperling, F. A. H. (2018). Would an RRS by any other name sound as RAD? *Methods in Ecology and Evolution*, *9*(9), 1920–1927.

Cock, P. J., Bonfield, A. J. K., Chevreux, B., & Li, H. (2015). *SAM/BAM format v1.5 extensions for de novo assemblies*. https://www.biorxiv.org/content/ https://doi.org/10.1101/020024v1

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., McVean, G., & Durbin, R. (2011). The variant call format and VCFtools. *Bioinformatics*, *27*(15), 2156–2158. https://doi.org/10.1093/bioinformatics/btr330

Davey, J. W., Hohenlohe, P. A., Etter, P. D., Boone, J. Q., Catchen, J. M., & Blaxter, M. L. (2011). Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature Reviews Genetics*, *12*, 499–510.

DePristo, M., Banks, E., Poplin, R., Garimella, K., Maguire, J., Hartl, C., Philippakis, A., del Angel, G., Rivas, M. A., Hanna, M., McKenna, A., Fennell, T., Kernytsky, A., Sivachenko, A., Cibulskis, K., Gabriel, S., Altshuler, D., & Daly, M. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, *43*, 491–498. https://doi.org/10.1038/ng.806

Doyle, J. J., & Doyle, J. L. (1987). A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochemical Bulletin*, *19*, 11–15.

Duche, P., & Salamin, N. (2020). A cautionary note on the use of genotype callers in phylogenomics. *Systematic Biology*. https://doi.org/10.1093/sysbio/syaa081

Eaton, D. A. R., Spriggs, E. L., Park, B., & Donoghue, M. J. (2017). Misconceptions on missing data in RAD-seq phylogenetics with a deep-scale example from flowering plants. *Systematic Biology*, *66*(3), 399–412.

Eguchi, K., Oguri, E., Sasaki, T., Matsuo, A., Nguyen, D. D., Jaitrong, W., Yahya, B. E., Chen, Z., Satria, R., Wang, W. Y., & Suyama, Y. (2020). Revisiting museum collections in the genomic era: Potential of MIG-seq for retrieving phylogenetic information from aged minute dry specimens of ants (Hymenoptera: Formicidae) and other small organisms. *Myrmecological News*, *30*, 151–169.

Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S., & Mitchell, S. E. (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS ONE*, *6*(5), e19379. https://doi.org/10.1371/journal.pone.0019379

Eriksson, C. E., Ruprecht, J., & Levi, T. (2020). More affordable and effective noninvasive single nucleotide polymorphism genotyping using high-throughput amplicon sequencing. *Molecular Ecology Resources*, *20*(6), 1505–1516.

Fama, N. M., Sinn, B. T., & Barrett, C. F. (2021). Integrating genetics, morphology, and fungal host specificity in conservation studies of a vulnerable, selfing, mycoheterotrophic orchid (*Corallorhiza bentleyi* Freudenst.). *Castanea*, *86*(1), 1–21.

Franchini, P., Parera, D. M., Kautt, A. F., & Meyer, A. (2017). quaddRAD: A new high-multiplexing and PCR duplicate remove ddRAD protocol produces novel evolutionary insights in a nonradiating cichlid lineage. *Molecular Ecology*, *26*(10), 2783–2795.

Glenn, T. C., Nilsen, R. A., Kieran, T. J., Sanders, J. G., Bayona-Vásquez, N. J., Finger, J. W., Pierson, T. W., Bentley, K. E., Hoffberg, S. L., Louha, S., & Garcia-De Leon, F. J. (2019). Adapterama I: Universal stubs and primers for 384 unique dual-indexed or 147,456 combinatorially-indexed Illumina libraries (iTru & iNext). *PeerJ*, *7*, e7755.

Goudet, J. (2005). Hierfstat, a package for R to compute and test hierarchical F-statistics. *Molecular Ecology Notes*, *5*(1), 184–186.

Gupta, M., Chyi, Y.-S., Romero-Severson, J., & Owen, J. L. (1994). Amplification of DNA markers from evolutionarily diverse genomes using single primers of simple-sequence repeats. *Theoretical and Applied Genetics*, *89*, 998–1006.

Gutiérrez-Ortega, J. S., Salinas-Rodríguez, M. M., Martínez, J. F., Molina-Freaner, F., Pérez-Farrera, M. A., Vovides, A. P., Matsuki, Y., Suyama, Y., Ohsawa, T. A., Watano, Y., & Kajita, T. (2018). The phylogeography of the cycad genus Dioon (Zamiaceae) clarifies its Cenozoic expansion and diversification in the Mexican transition zone. *Annals of Botany*, *121*(3), 535–548.

Jackman, S. D., Vandervalk, B. P., Mohamadi, H., Chu, J., Yeo, S., Hammond, S. A., Jahesh, G., Khan, H., Coombe, L., Warren, R. L., & Birol, I. (2017). ABySS 2.0: Resource-efficient assembly of large genomes using a Bloom filter. *Genome Research*, *27*(5), 768–777.

Jombart, T., & Ahmed, I. (2011). adegenet 1.3-1: New tools for the analysis of genome-wide SNP data. *Bioinformatics*, https://doi.org/10.1093/bioinformatics/btr521

Kamvar, Z. N., Brooks, J. C., & Grünwald, N. J. (2015). Novel R tools for analysis of genome-wide population genetic data with emphasis on clonality. *Frontiers in Genetics*, *6*, 208.

Lemmon, E. M., & Lemmon, A. R. (2013). High-throughput genomic data in systematics and phylogenetics. *Annual Review of Ecology, Evolution, and Systematics*, *44*, 99–121.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., & 1000 Genome Project Data Processing Subgroup. (2009). The Sequence alignment/map (SAM) format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079.

Macaya-Sanz, D., Chen, J., Kalluri, U. C., Muchero, W., Tschaplinski, T. J., Guner, L. E., Simon, S. J., Biswal, A. K., Bryan, A. C., Payyavula, R., Xie, M., Yang, Y., Zhang, J., Mohnen, D., Tuskan, G. A., & DiFazio, S. P. (2017). Agronomic performance of *Populus deltoides* trees engineered for biofuel production. *Biotechnology for Biofuels*, *10*, 253.

Marandel, F., Charrier, G., Lamy, J.-B., Le Cam, S., Lorance, P., & Trenkel, V. M. (2020). Estimating effective population size using RADseq: Effects of SNP selection and sample size. *Ecology and Evolution*, *10*, 1929–1937.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., & DePristo, M. A. (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, *20*, 1297–1303. https://doi.org/10.1101/gr.107524.110

Meek, M. H., & Larson, W. A. (2019). The future is now: Amplicon sequencing and sequence capture usher in the conservation genomics era. *Molecular Ecology Resources*, *19*(4), 795–803.

Narum, S. R., Buerkle, C. A., Davey, J. W., Miller, M. R., & Hohenlohe, P. A. (2013). Genotyping-by-sequencing in ecological and conservation genomics. *Molecular Ecology*, *22*(11), 2841–2847.

Ortiz, E. M. (2019). *vcf2phylip v2.0: convert a VCF matrix into several matrix formats for phylogenetic analysis*. https://doi.org/10.5281/zenodo.2540861

Park, J. S., Takayama, K., Suyama, Y., & Choi, B.-H. (2019). Distinct phylogeographic structure of the halophyte Suaeda malacosperma (Chenopodiaceae/Amaranthaceae), endemic to Korea-Japan region, influenced by historical range shift dynamics. *Plant Systematics and Evolution*, *305*, 193–203. https://doi.org/10.1007/s00606-018-1562-8

Peterson, B. K., Weber, J. N., Kay, E. H., Fisher, H. S., & Hoekstra, H. E. (2012). Double digest RADseq: An inexpensive method for *de novo* SNP discovery and genotyping in model and non-model species. *PLoS One*, *7*(5), e37135.

Piwczyński, M., Trzeciak, P., Popa, M.-O., Pabijan, M., Corral, J. M., Spalik, K., & Grzywacz, A. (2020). Using RAD seq for reconstructing phylogenies of highly diverged taxa: A test using the tribe Scandiaceae (Apiaceae). *Journal of Systematics and Evolution*, *59*(1), 58–72.

Poplin, R., Ruano-Rubio, V., DePristo, M. A., Fennell, T. J., Carneiro, M. O., Van der Auwera, G. A., Kling, D. E., Gauthier, L. D., Levy-Moonshine, A., Roazen, D., Shakir, K., Thibault, J., Chandran, S., Whelan, C., Lek, M., Gabriel, S., Daly, M. J., Neale, B., MacArthur, D. G., & Banks, E. (2017). Scaling accurate genetic variant discovery to tens of thousands of samples. *bioRxiv* preprint. https://doi.org/10.1101/201178

R Core Team. (2019). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org

Sinn, B. T., Simon, S. J., & Barrett, C. F. (2021). *ISSRseq Pipeline Release v1.0.0*. https://zenodo.org/badge/latestdoi/189902579

Suyama, Y., & Matsuki, Y. (2015). MIG-seq: An effective PCR-based method for genome-wide single-nucleotide polymorphism genotyping using the next-generation sequencing platform. *Scientific Reports*, *5*, 16963.

Takata, K., Taninaka, H., Nonaka, M., Iwase, F., Kikuchi, T., Suyama, Y., Nagai, S., & Yasuda, N. (2019). Multiplexed ISSR genotyping by sequencing distinguishes two precious coral species (Anthozoa: Octocorallia: Coralliidae) that share a mitochondrial haplotype. *PeerJ*, *7*, e7769. https://doi.org/10.7717/peerj.7769

Tamaki, I., Yoichi, W., Matsuki, Y., Suyama, Y., & Mizuno, M. (2017). Inconsistency between morphological traits and ancestry of individuals in the hybrid zone between two *Rhododendron japonoheptamerum* varieties revealed by a genotyping-by-sequencing approach. *Tree Genetics and Genomes*, *13*(1), 4. https://doi.org/10.1007/s11295-016-1084-x

Tripp, E. A., Tsai, Y.-H.-E., Zhuang, Y., & Dexter, K. G. (2017). RADseq dataset with 90% missing data fully resolves recent radiation of *Petalidium* (Acanthaceea) in the ultra-arid deserts of Namibia. *Ecology and Evolution*, *7*, 7920–7936.

Van Tassell, C. P., Smith, T. P. L., Matukumalli, L. K., Taylor, J. F., Schnabel, R. D., Lawley, C. T., Haudenschild, C. D., Moore, S. S., Warren, W. C., & Sonstegard, T. S. (2008). SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. *Nature Methods*, *5*(3), 247–252.

Zietkiewicz, E., Rafalski, A., & Labuda, D. (1994). Genome fingerprinting by Simple Sequence Repeat (SSR)-anchored polymerase chain reaction amplification. *Genomics*, *20*, 176–183.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Sinn, B. T., Simon, S. J., Santee, M. V., DiFazio, S. P., Fama, N. M., & Barrett, C. F. (2022). ISSRseq: An extensible method for reduced representation sequencing. *Methods in Ecology and Evolution*, 13, 668–681. https://doi.org/10.1111/2041-210X.13784