Simple sequence repeats (SSRs), also known as microsatellites or simple tandem repeats, are common genetic markers used for genetic analyses (Hodel et al., 2016). Their defining characteristic is a short (di-, tri-, or tetranucleotide) DNA sequence repeated multiple times. The mutation rate for SSRs, attributed to polymerase slippage during DNA replication and repair, is substantially higher than that for substitution and indel-type mutations, making them multi-allelic and highly informative. SSRs are generally considered selectively neutral, being influenced mainly by gene flow, genetic drift, and mutation, making them particularly useful for population genetic studies (Slatkin, 1995; Spencer et al., 2000; Nielsen, 2005). SSRs have also been utilized extensively for the identification of quantitative trait loci, especially when combined with linkage map construction (Fang et al., 2016; Guo et al., 2017). Traditionally, SSR genotypes were determined using the PCR and primers from conserved flanking regions of the SSR motif, followed by staining with an intercalating agent, often ethidium bromide. Both polyacrylamide and ethidium bromide pose human health risks. Alternatively, SSR genotypes can be determined from fluorophore-labeled PCR products using specialized fragment analysis capillary sequencing equipment. Although fluorophore-labeled primers are safer, they are significantly more expensive and require substantial capital equipment investments. Fluorophore-labeled SSRs can be multiplexed to reduce cost; however, multiplex schemes require a significant investment of time to develop and are limited in capacity, allowing typically fewer than nine SSRs (e.g., three PCR product sizes and three fluorophore types). Furthermore, SSR genotyping is complicated by the inherent lab-to-lab variation due to protocol and instrumentation differences, primer sequence changes, and the lack of a standardized calling method, which all hinder the ability to carry data from one experiment to the next.

The advent of next-generation sequencing dramatically changed the way putative SSR loci were identified. Traditionally, putative SSR markers were identified from sequenced genomic DNA libraries that were enriched for SSR motifs using oligonucleotide probe sequences complementary to a targeted SSR motif. The development of whole-genome assemblies allowed for the application of SSR genotyping to en masse genotyping from resequencing data for diversity panels and linkage mapping populations. SSRgenotyper extends SSR identification to en masse genotyping from resequencing data for diversity panels and linkage mapping populations.

**PREMISE:** Many programs can identify simple sequence repeat (SSR) motifs in genomic data. SSRgenotyper extends SSR identification to en masse genotyping from resequencing data for diversity panels and linkage mapping populations.

**METHODS AND RESULTS:** SSRgenotyper will find and genotype SSRs from SAM files and an SSR reference FASTA. Several outputs are possible, including a simple table with the SSR marker name, position, and SSR alleles, defined by the repeat number of the repeat motif. Specific output files include a GENEPOP-formatted file for downstream genetic diversity analyses and a traditional A, H, B mapping file output that is phased to the parents of the population for biparental linkage map construction. Linkage maps produced using SSRgenotyper genotypes were highly collinear with physical maps and correctly inferred known phylogenies.

**CONCLUSIONS:** SSRgenotyper provides an easy-to-use, accurate, and scalable SSR genotyping platform for whole-genome resequencing data. SSRgenotyper is freely available at https://github.com/dlewis27/SSRgenotyper.

**KEYWORDS:** genotyping; linkage mapping; microsatellite; simple sequence repeat; simple tandem repeat.
bioinformatic approaches for the identification of SSR motifs directly from sequencing data. Programs like MISA (Beier et al., 2017), Krait (Du et al., 2018), and Kmer-SSR (Pickett et al., 2017), among others, can quickly and accurately identify thousands of SSR loci targets within a genome, thus providing researchers with thousands of SSR loci that can be targeted for marker development. Advances in sequencing technology continue to drive down the cost of sequencing, making SSR genotyping-by-sequencing on whole populations an attractive alternative to the traditional PCR/electrophoresis-based method. Sequencing costs are now within a range that resequencing even non-model populations, especially species with small to intermediate-sized genomes, is affordable. We report SSRgenotyper, an SSR genotyping pipeline that is fast and accurate and requires only resequencing data (whole genome or reduced representational), thereby avoiding many of the potential difficulties associated with the traditional methods for genotyping SSR loci.

METHODS AND RESULTS

SSRgenotyper directly calls SSR genotypes from sequencing reads for individuals within natural populations, germplasm collections, or segregating biparental mapping populations. The program will find di-, tri-, and tetranucleotide-based SSRs from Sequence Alignment/Map (SAM) files and an SSR reference FASTA file. Mononucleotide SSRs are excluded from the analysis as are soft-masked sequences in the SSR reference FASTA file. SSRgenotyper has only been tested on diploid and allopolyploid organisms that show normal disomic inheritance. Because the maximum number of alleles detected at an SSR locus by SSRgenotyper is currently limited to two, SSRgenotyper should be used with caution on allopolyploids that show non-disomic inheritance (e.g., trisomy, tetrasomy, etc.) and autopolyploids. An outline of the workflow is provided in Figure 1.

SSR reference FASTA

SSRgenotyper requires an SSR reference FASTA file that provides sequence information for each targeted SSR to be genotyped, including 100 bp of upstream and downstream flanking sequence. The SSR reference FASTA file can contain SSR sequence information derived a priori from previous SSR discovery projects or as a comprehensive representation of all SSRs found within the targeted species (often tens of thousands of putative SSR loci). If a genome assembly is available for the target species, the development of a comprehensive SSR reference FASTA file can be easily created using MISA, BEDTools (Quinlan and Hall, 2010), and several shell commands. Step-by-step directions for producing a comprehensive SSR reference FASTA file from a reference genome are provided in Appendix 1. In short, MISA is used to produce a general feature format (GFF) file containing the locations of the SSR loci within the reference genome assembly. The reference genome can be a gold standard, chromosomal-scale reference genome or a simple draft assembly consisting of thousands of contigs. Shell commands are used to remove compound SSRs and any loci with insufficient flanking sequence (<100 bp; those at the edges of contigs). BEDTools is then used to extract the sequences of the SSR loci, including 100 bp of flanking sequence, to make the comprehensive SSR reference FASTA file.

FIGURE 1. SSRgenotyper workflow. Sequences from each individual are mapped against a FASTA file containing the targeted SSR and 100 bp of flanking sequence. SSR targets can be developed from reference genomes using SSR-finding software (Appendix 1) or as predeveloped SSRs. Read mapping, PCR duplicate removal, and read mapping quality control can be accomplished with a variety of alignment softwares (Appendix 2). SAM files are then utilized by SSRgenotyper for genotyping. Alleles are called based on their repeat number. Output files include GENEPOP and linkage analysis formats.
Read mapping

Sequencing reads for each individual in the population are mapped to the SSR reference FASTA file to generate a SAM file for each individual. SSRgenotyper has no requirement for a specific read type (technology or length), only requiring that the reads should be highly accurate to maintain genotyping quality. Reads should be quality controlled and trimmed (e.g., Trimomatic [Bolger et al., 2014]) prior to mapping (e.g., BWA-MEM [Li, 2013], minimap2 [Li, 2018], and bowtie2 [Langmead and Salzberg, 2012]) to the SSR reference FASTA file. Because SSRgenotyper makes allelic calls based on ratios of alleles identified from mapped reads, it is important to remove PCR duplicate reads and any poorly mapped reads from the SAM files. We provide step-by-step details in Appendix 2 for using BWA-MEM for read mapping of Illumina paired-end reads and processing the resulting SAM files, including simple loop commands for handling multiple individuals in large populations.

SSRgenotyper

SSRgenotyper, coded in Python 3, requires three positional arguments, specifically an SSR reference FASTA file (Appendix 1), a text file listing names of the SAM files for each individual in the population (Appendix 2, Step 5), and an output name for the results. SSRgenotyper (1) identifies the boundaries of the SSR repeat using the boundary nucleotides specified by the user, (2) verifies that the SSR is a perfect match to the reference repeat (removing any reads with mismatches; only perfect repeats are analyzed), (3) calculates the total nucleotide length of the repeat, and (4) determines the allele call by dividing the repeat length by nucleotide motif type (two for di-, three for tri-, or four for tetranucleotide repeats). Alleles are thus based on the number of repeats rather than on PCR product sizes, making them more easily transferable across experiments and laboratories. Processing of the SAM files is fast with minimal memory utilization and can be accomplished on an inexpensive desktop computer. Several optional arguments provide for fine tuning of the genotype-calling algorithm, including:

- **Support.** The minimum number of supporting reads needed for a genotype to be called at an SSR locus. If the total number of reads is less than this threshold, the genotype will be reported as missing. For example, if six reads are required and there are three reads supporting the major allele and two reads supporting the minor allele, no alleles will be called as the minimum threshold of six total reads was not met.

- **MinorAlleleHet.** The minimum percentage threshold of reads required for the minor allele for a genotype to be considered heterozygous. For diploid organisms, a maximum of two alleles should be present at the SSR locus (a minor and major allele). This parameter sets the cut-off value for calling an SSR as a heterozygote. If two alleles are found but the percentage of minor allele reads is below this cutoff, the genotype will be called as a homozygote for the major allele. If the species or population is expected to be homozygous (i.e., cleistogamous, recombinant inbred lines, etc.), setting this argument to 0.51 will produce only homozygous genotypic calls.

- **BorderSeq.** The number of nucleotides on each side of the reference SSR motif that will be evaluated to determine whether a read supports a genotypic call. If there is a concern with nonspecific read mapping (e.g., allopolyplody, paralogous/duplicated loci), increasing this parameter can add specificity to the allele calling. See also --mismatch.

- **spuriousAlleleRemoval.** PCR amplification of SSR loci is known to produce spurious (stutter) alleles that can result in ambiguous genotypes. These spurious alleles are normally minimal in frequency but need to be removed for accurate genotyping. Because diploid species should only have a maximum of two alleles, SSR loci that have more than the expected two (major/minor) alleles are marked as ambiguous (missing) if the additional allele(s) is supported by more than this threshold percentage of reads; otherwise, the allele(s) is considered to be spurious and is excluded.

- **mismatch.** The number of mismatches (substitutions and indels) allowed in the border sequence (--BorderSeq) region between the reads and the reference sequence. For more distantly related individuals (e.g., different species or subspecies), the number can be increased to accommodate expected sequence differences.

SSRgenotyper also provides several data filtering options. The filtering parameters set missing data thresholds for removing SSR loci and/or individuals from the final output results files based on the percentage of missing data:

- **FilterDataLoci.** The maximum missing data threshold for reporting an individual (SAM file). For example, if “--FilterDataLoci 0.3” then any SSR locus with more than 30% missing data is removed from the output files.

- **FilterDataSam.** The maximum missing data threshold for reporting an individual (SAM file). For example, if “--FilterDataSam 0.3” then any individual with more than 30% missing data is removed from the output files.

Several result (output) files are produced by SSRgenotyper, including:

- **The .ssr file.** The .ssr file is a basic genotyping results file in a tab-delimited table with SSR loci in rows and individuals (SAM files) in columns. The genotypes are called as alleles that reflect the number of repeat units found at each SSR locus. For example, a “9,9” genotypic call reflects a homozygote. Similarly, a “9,7” genotypic call reflects a heterozygote where reads with repeat numbers of “9” or “7” were identified. Any genotypes that are coded as “missing” in this file are coded to provide additional information as to why the genotype was coded as missing. Specifically, missing data are coded as “0,1” (i.e., the reference SSR was not evaluated by SSRgenotyper for one of several possible reasons, including the reference sequence was soft masked or the repeat was not a di-, tri-, or tetranucleotide repeat, see also --RefUnitsMin), “0,2” (no reads mapped to the SSR locus), “0,3” (more than the expected two alleles were identified in the mapped reads, see also --spuriousAlleleRemoval), or “0,4” (insufficient reads were identified to support calling a genotype, see also --Support).

- **The .ssrstat file.** This file includes basic run and genotyping statistics, including a listing of the positional and optional arguments selected as well as various run statistics. The statistics calculated for each of the various output file types (.ssr, .pop, etc.).
.map) take into account the data filtering arguments. The statistics reported include total number of individuals reported, total number of SSR loci reported, total number of genotypes called (separated as homozygous and heterozygous calls), and the percentage of missing calls.

- **The .pop file.** Eliciting the --Genepop argument produces a GENEPOP-formatted file. GENEPOP (Rousset, 2008) is a popular population genetics program that can be used to test for Hardy–Weinberg exact tests, linkage disequilibrium, population differentiation, and gene flow estimates. GENEPOP files can be easily converted to other file formats used by other population genetics platforms (e.g., ARLEQUIN [Excoffier and Lischer, 2010], STRUCTURE [Falush et al., 2007], MSA [Dieringer and Schlatterer, 2003]) using PGDSpider (Lischer and Excoffier, 2012). All missing data are coded as "000000". Monomorphic loci are excluded in the output.

- **The .map file.** For segregating mapping populations, eliciting the --LinkageMapFile argument produces an "A,B,H" genotype file to facilitate downstream linkage map construction of biparental populations using linkage map construction software (e.g., JoinMap [Van Ooijen, 2006]). The .map file is a tab-delimited table with SSR loci in rows and individuals in columns. The first two SAM files in the SAMFiles.txt (Appendix 2, Step 5) should correspond to the two parents of the biparental population and are necessary for correctly phasing the genotypes. By default, if the parental genotypes are monomorphic or if one of the parents is missing a genotypic call, the SSR locus is excluded from the output. If one or both of the parents have a heterozygous genotype, phase cannot be determined, and the marker is excluded. Similarly, any progeny genotypic calls that do not agree with the alleles identified in the parents are deemed problematic and are assigned a missing value ("."). SSRgenotyper can impute the genotype of a missing parent (the other parent must have a genotype) based on the progeny by providing the argument with a specific numeric threshold (i.e., --LinkageMapFile 0.30). If a threshold is provided and two alleles (major and minor) are found in the progeny with the minor allele frequency above the given threshold and where one allele agrees with the genotyped parent, SSRgenotyper will internally assign the second allele as the genotype of the missing parent; however, the imputed genotype will still be reported in the .map file as missing, thus allowing the user to identify which markers had an imputed parental genotype.

- **The Alignment file.** Read alignments for a specific individual and a specific SSR locus can be elicited using the --AlignmentShow argument. Eliciting this argument will produce a .txt file showing the read alignments for the specific individual to the reference sequence. This file is particularly useful for spot checking alignments and specific genotypic calls.

A complete description, including defaults, of all optional arguments and output file types is provided in the README document on https://github.com/dlewis27/SSRgenotyper.

**Accuracy and error profile**

To assess the accuracy of SSRgenotyper, we evaluated the concordance of genotype calling from replicated Illumina paired-end (2 × 100 bp) sequencing runs of the quinoa (Chenopodium quinoa Willd.) accession PI 614886. The full quinoa genome was previously reported by Jarvis et al. (2017). Quinoa is an allopolyploid (2n = 4x = 36) diot that is generally considered to be cleistogamous (self-pollinating) and thus exhibits low levels of heterozygosity. The sequencing runs were derived from the same sequencing library (same plant and DNA extraction), each of which achieved ~28× genome coverage (BioProject no.: PRJNA640095). Varying only the number of reads necessary to support a genotype (--Support 1 to 10; full command provided in Appendix 3), SSRgenotyper genotyped between 40,301 and 99,361 SSR loci (Fig. 2A), of which an average of 97.5% of the genotypic calls were concordant. Not unexpectedly, as more read support was required, the concordance between the data set increased from 97.0% to 97.8%, reflecting that as more reads were available for specific SSR loci, more accurate calling of the SSR was produced. Analysis of the discordant calls (at --Support = 6; 56,477 SSR loci) suggested that motif size (di-, tri-, and tetranucleotide) was an important factor in calling accuracy, with discordance being highest with dinucleotide and tetranucleotide motifs (3.3% and 3.2%, respectively) compared to trinucleotide motifs (1.7%; Fig. 2B). Trinucleotide-motif SSR were the most common SSR identified in the data, representing 61.3% of all SSRs genotyped, whereas dinucleotide-motif and tetranucleotide-motif SSRs represented 31.3% and 7.4% of all SSRs, respectively. The most discordant dinucleotide motif, AT/AT, showed an error rate of 7.9%, while all other complementary motifs (AG/CT, AC/GT) had discordance rates of 2.6% (the complementary motif CG/CG was excluded from analysis as it represented fewer than 0.5% of all dinucleotide SSRs). A positive correlation (R² = 0.79) was also observed between discordance and the initial SSR length as identified in the reference sequence, with an average discordance of 2.1% for SSRs that exhibited 4–9 repeats in the reference sequence and 3.2% for SSRs with 10–15 repeats (Fig. 2C). Although very few SSRs (0.27%) were identified with >15 repeats in the reference sequence, they showed the highest discordance at 13.8%.

In a second analysis, we evaluated the concordance between repeat number reported in a reference genome for Avena atlantica B. R. Baum & Fedak (BioProject no.: PRJNAS46592) with the SSR repeat number called directly (de novo) from the Illumina raw data using SSRgenotyper (--Support = 5; 96,833 SSR loci). Avena atlantica is a diploid monocot species (2n = 2x = 14) with a highly repetitive genome (Maughan et al., 2019). The results between the two analyses are similar, with discordance being highest with dinucleotide and tetranucleotide motifs (5.7% and 5.0%, respectively) compared to trinucleotide motifs (1.2%; Appendix S1). As in the first analysis, trinucleotide-motif SSRs were the most common SSR identified in the data, representing 72.9% of all SSR loci genotyped. The most discordant dinucleotide motif was also the AT/AT motif, which showed an error rate of 13.4%, while all other complementary motifs (AG/CT, AC/GT, CG/CG) had much lower discordant rates of 4.8%, 4.2%, and 2.6%, respectively. Similarly, a positive correlation (R² = 0.668) was also observed between discordance and the SSR length identified in the reference sequence, with a discordance of 2.4% for SSRs that exhibited 4–9 repeats, which represented 97.8% of all SSRs in the data set. Approximately 1.8% of the data set was represented by SSRs with 10–15 repeats, which showed a 4.7% discordance. SSRs with greater than 15 repeats were the most discordant (20.3%) but made up less than 0.4% of the total number of the SSRs.

Increasingly large repeat numbers and increased AT content are known to cause slippage problems (stuttering) for Taq DNA polymerase (Brookes et al., 2012; Hosseinzadeh-Colagari et al., 2016), as
well as increased difficulty for read-mapping algorithms. Taken together, these results suggest that increased coverage combined with focused analysis of trinucleotide repeats (or removal of problematic AT/AT dinucleotide-motif types) with starting reference SSR repeats of sizes <10 will produce the most accurate genotyping data sets.

Implementation

To demonstrate the utility of SSRgenotyper and to further validate the genotyping, we generated several publicly available resequencing data sets, including two segregating mapping populations that used different sequencing strategies (RNA-seq and whole-genome resequencing) and a genetic diversity panel. For the segregating populations, SSRgenotyper was used to produce phased linkage mapping data sets (--LinkageMapFile) that were used to construct genetic linkage maps. The SSR linkage maps were then compared to published chromosome-scale genome assemblies.

The first segregating data set was a reduced representational RNA-seq experiment consisting of 155 recombinant inbred lines (RIL) from the cross of Arabidopsis thaliana Bayreuth-0 x Shahdara (Serin et al., 2017) (BioProject no.: PRJNA418075). HISAT2 (Kim et al., 2019), a splice-aware read mapper, was used to map the RNA-seq reads to the SSR reference FASTA file generated from the TAIR9 genome A. thaliana assembly (Lamesch et al., 2012) as detailed in Appendix 1. Putative SSR loci identified in the TAIR9 assembly were named according to chromosomal origin and position to allow downstream comparisons between the physical genome and the constructed linkage map. As this is an RIL population, SSRgenotyper was run with the --MinorAlleleHet argument set to 0.51 to produce only homozygous genotypes. The complete SSRgenotyper command, including all arguments, is provided in Appendix 3.

SSRgenotyper reported a total of 51,993 genotypes from 327 SSR loci across 159 individuals (including the two parents) in ~7 min using 279 MB of memory, of which 5355 (10.3%) were missing genotypes. As expected for an RIL population, where a 1:1 (A:B) segregation ratio is expected, 22,791 (49.6%) and 23,193 (50.4%) A (Bayreuth-0) and B (Shahdara) genotypes were observed, respectively. Eighteen loci were subsequently dropped from the linkage map construction due to significant segregation distortion (P value < 1E-08). Severely distorted markers are a natural occurring phenomenon and are commonly excluded from linkage map construction as they obscure linkage relations and distort linkage distances (Zuo et al., 2019). JoinMap 5 (Van Ooijen, 2006) grouped the remaining 309 SSR loci into five linkage groups based on the independence logarithm of the odds scores (LOD ≥ 5.0), calculated for recombination frequency using the G2 statistic. SRs within groups were then ordered using regression mapping and corrected with the Kosambi mapping function to determine centimorgan (cM) distances. A comparison of the linkage maps with the complete A. thaliana genome assembly clearly identified a one-to-one association between the ordered linkage groups and the A. thaliana chromosomes (n = 5). Without exception, each SSR locus within a linkage group originated from the same physical chromosome. Moreover, the collinearity of the linkage maps with the physical genome was highly correlated (R² > 0.98), suggesting accurate genotyping of the SSRs (Appendix S2). Several of the linkage groups exhibited areas, corresponding with the physical locations of the centromere (Hosouchi et al., 2002), where the density of SSR loci was noticeably deficient (e.g., Chromosome 2). Considering that these SSRs were derived from an RNA-seq data set, this observation was not

FIGURE 2. Accuracy and error profile for SSRgenotyper. (A) Total number of genotypes reported that were concordant (blue) and discordant (red) at various minimum read support levels. (B) Percent concordant (blue) and discordant (red) genotypes based on di-, tri-, or tetranucleotide motif type. (C) Percent concordant (blue) and discordant (red) genotypes based on repeat length as identified in the reference FASTA file.
unexpected given that centromeric regions are well documented to have markedly decreased concentrations of genic sequences.

The second segregating data set tested was a whole-genome resequencing experiment in foxtail millet (Setaria italica (L.) P. Beauvois) consisting of 120 RILs from a cross between Longgu7 × Yugu1 (Liu et al., 2020) (BioProject no.: PRJNA562988). BWA-MEM was used to map the whole-genome resequencing reads to the SSR reference FASTA file generated from the S. italica version 2.2 genome assembly (Bennetzen et al., 2012). The complete SSRgenotyper command, including all arguments, is provided in Appendix 3. SSRgenotyper reported 102,720 genotypes, including 49,762 (54.3%) and 41,846 (45.7%) phased as the A (Longgu7) and B (Yugu1) genotype, respectively, from 858 SSR loci across 120 segregating individuals in ~46 min using 2 GB of memory. Slightly less than 11% (11,112) of the genotypes were assigned as missing. Thirty-two loci were removed for significant segregation distortion (P value < 1E-09). Liu et al. (2020) previously attributed the segregation distortion observed in this population to various factors, including potential partial sterility and/or gametic/zygotic selection. The remaining 819 SSR loci were grouped (LOD ≥ 4.0) into 11 linkage groups and ordered using the regression mapping in JoinMap5, corrected with the Kosambi mapping function. All SSRs within a group originated from a single chromosomal scaffold, including three small linkage groups that mapped to S. italica scaffold 2; the remaining eight, much larger linkage groups corresponded with each of the remaining eight S. italica (n = 9) chromosomal scaffolds. Although the overall linkage and physical maps were highly collinear (Appendix S3), nearly all of the linkage groups exhibited an area of concentrated linkage distance (cM) compression relative to the physical map distance. These compressed areas are presumed to correspond to the physical location of the centromere. Crossing over is well known to be suppressed at centromeres, ranging from fivefold to >200-fold depending on the organism (Haupt et al., 2001; Talbert and Henikoff, 2010). Thus, the centromeres, which often span several hundreds of megabases of physical space, appear compacted in the linkage groups comparisons. The lack of recombination events in the centromere makes accurately ordering the loci within the centromere significantly more difficult; nonetheless, the observed collinearity and accurate prediction of the centromeric region are suggestive of accurate genotyping.

For the genetic diversity panel analysis, SSRgenotyper was used to produce a genetic diversity data set (–Genepop) from which phylogenies were constructed and compared for congruency to published single-nucleotide polymorphism (SNP)–based phylogenies constructed from the same resequencing data. We utilized a whole-genome resequencing panel reported by Jarvis et al. (2017) (PRJNA306026), representing 15 quinoa (Chenopodium quinoa) accessions representing the two major ecotypes of quinoa: highland and coastal. Also included were five accessions of North American C. berlandieri Moq. (var. sinutatum (Murr) Wahl, subsp. nuttalliae (Saff.) H. Dan. Wilson & Heiser [huazontle], var. macrocalycium (Aellen) Cronquist, var. boscianum (Moq.) Wahl and var. zschackei (Murr) Murr ex Graebn.) and one accession each of C. berlandieri sinuatum (Murr) Wahl, subsp. Moq. (var. accession representing the two major ecotypes of quinoa: highland). Chenopodium quinoa is believed to have been domesticated from wild C. hircinum, whereas C. berlandieri is highly diverse and basal to the tetraploid chenopod complex. Using a minimum read support of four and a minor allele rate of 0.25 for calling heterozygotes (Appendix 3), SSRgenotyper reported 165,157 genotypes, including 145,176 (88%) homozygous and 19,981 (12%) heterozygous genotypes from 9450 SSR loci across the 22 Chenopodium accessions in ~35 min using 5.4 GB of memory. Populations version 1.2.31 (Langella, 1999) was used to generate a Nei’s standard (Dh) genetic distance matrix from the GENEPOP file, which was subjected to cluster analysis to produce a NeighborNet phylogeny that was visualized in splitstree4 (Huson, 1998; Bryant and Moulton, 2002; Kloepfer and Huson, 2008). Identical to the SNP phylogeny, our SSR phylogeny clearly identifies three clades, one for each of the quinoa ecotypes and one consisting of the basal C. berlandieri species. Moreover, our SSR analysis matches the SNP phylogeny placement of one of the C. hircinum accessions (BYU 566) at a basal position to the coastal ecotypes, with the other C. hircinum basal to the C. quinoa clade (Appendix S4). Jarvis et al. (2017) hypothesized that this placement of C. hircinum suggests the possibility that the different quinoa ecotypes were derived from potentially multiple independent domestication events of wild C. hircinum. Another explanation could be crop–weed gene flow between geographically related C. quinoa and C. hircinum accessions, which are known to form fertile hybrids (Wilson, 1990).

CONCLUSIONS

SSRgenotyper provides an easy-to-use, accurate, and scalable SSR genotyping platform for resequencing data (whole-genome or reduced representational). Traditional SSR experiments include an SSR discovery phase typically using SSR-enriched libraries and oligo primer design, testing for amplification and informativeness on testing panels. All of these steps are circumvented using resequencing data, as any resequencing analysis will include hundreds of informative SSRs. SSRgenotyper reports SSR genotypes as repeat numbers, which are easily transferred across experiments and laboratories, thereby providing a universal standard that has been missing for SSR genotyping. While the accuracy of the genotyping was shown to be high (>97%), improvements in genotyping accuracy can be achieved through increased sequencing coverage and a focus on trinucleotide-motif SSRs as they showed the lowest error rate in our analysis. The continued decrease in costs associated with next-generation sequencing should make SSR genotyping-by-sequencing increasingly affordable, even for non-model species and especially for small-genome species. Perhaps the most problematic aspect of SSR genotyping is the amplification slippage that occurs during the PCR process which produces aberrant “stutter” products. These stutter products can obscure the true alleles. Generally, the frequency of the stutter products is minimal and they can be removed by SSRgenotyper using filters against unexpected low-frequency alleles (i.e., --spuriousAlleleRemoval). Furthermore, we hypothesize that spurious alleles can be further decreased using PCR-free sequencing libraries (e.g., Illumina’s TruSeq PCR-Free library kit) or other emerging next-generation sequencing methods that avoid PCR, such as Pacific Biosciences high-fidelity single-molecule sequencing technologies.

ACKNOWLEDGMENTS

The authors thank Michael Greenburg (Office of Research Computing, Brigham Young University, Provo, Utah) for his expert help regarding high-performance computing.
AUTHOR CONTRIBUTIONS

PJ.M. and D.E.J. conceived and designed the program. D.H.L. wrote the Python code. PJ.M. and D.H.L. tested the program with the various test data sets. PJ.M. and D.E.J. wrote the manuscript. All authors read and approved the final manuscript.

DATA AVAILABILITY

SSRgenotyper is freely available at https://github.com/dlewis27/SSRgenotyper. Several test data sets were utilized in this project which are publicly available in GenBank, including Chenopodium quinoa (PRJNA640095, SRR12041554–SRR12041555, PRJNA306026, SRR4300210–SRR4300229), Arabidopsis thaliana (PRJNA418075, SRR6292944–SRR6293128), and Setaria italica (PRJNA562988, SRR10038747–SRR10038795).

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

APPENDIX S1. Accuracy and error profile for SSRgenotyper for Avena atlantica. (A) Total number of concordant (blue) and discordant (orange) genotypes based on repeat length as identified in the reference FASTA file. The y-axis scale is log10. (B) Total number of concordant (blue) and discordant (red) genotypes based on di-, tri-, or tetranucleotide motif type. Numbers for each class are indicated.

APPENDIX S2. Collinearity analysis of the Arabidopsis thaliana physical maps (in mega base pairs [Mbp]) and the linkage maps (in centimorgans [cM]) based on SSR genotyping. Each dot represents a genotyped SSR locus. A dotted oval identifies the predicted region of the centromere (Hosouchi et al., 2002).

APPENDIX S3. Collinearity analysis of the Setaria italica physical maps (in mega base pairs [Mbp]) and the linkage maps (in centimorgans [cM]) based on SSR genotyping. Each dot represents a genotyped SSR locus. The number of SSR loci mapped to each chromosomal scaffold is provided.

APPENDIX S4. Evolutionary relationships of Chenopodium species using a NeighborNet phylogenetic network generated from 9450 SSR loci. Red and blue lines identify coastal and highland C. quinoa ecotypes, while green and black lines identify C. berlandieri and C. hircinum accessions, respectively.

LITERATURE CITED

Beier, S., T. Thiel, T. Munch, U. Scholz, and M. Mascher. 2017. MISA-web: A web server for microsatellite prediction. Bioinformatics 33: 2583–2585.
Bennetzen, J. L., J. Schmutz, H. Wang, R. Percifield, J. Hawkins, A. C. Pontaroli, M. Estep, et al. 2012. Reference genome sequence of the model plant Setaria. Nature Biotechnology 30: 555–561.
Bolger, A. M., M. Lohse, and B. Usadel. 2014. Trimmomatic: A flexible trimmer for Illumina sequence data. Bioinformatics 30: 2114–2120.
Brookes, C., J. A. Bright, S. Harbison, and J. Buckleton. 2012. Characterising stutter in forensic STR multiplexes. Forensic Science International: Genetics 6: 58–63.
Bryant, D., and V. Moulton. 2002. NeighborNet: An agglomerative method for the construction of planar phylogenetic networks. In R. Guigó and D. Gusfield [eds.], Algorithms in bioinformatics. Lecture Notes in Computer Science, vol. 2452, 375–391. Springer, Berlin, Germany.
Dieringer, D., and C. Schlötterer. 2003. MICROSATellite ANALYSer (MSA): A platform independent analysis tool for large microsatellite data sets. Molecular Ecology Notes 3: 167–169.
Du, L. M., C. Zhang, Q. Liu, X. Y. Zhang, and B. S. Yue. 2018. Krait: An ultra-fast tool for genome-wide survey of microsatellites and primer design. Bioinformatics 34: 681–683.
Excoffier, L., and H. E. L. Lischer. 2010. Arlequin suite ver 3.5: A new series of programs to perform population genetics analyses under Linux and Windows. Molecular Ecology Resources 10: 564–567.
Falush, D., M. Stephens, and J. K. Pritchard. 2007. Inference of population structure using multilocus genotype data: Dominant markers and null alleles. Molecular Ecology Notes 7: 574–578.
Fang, X. M., K. J. Dong, X. Q. Wang, T. P. Liu, J. H. He, R. Y. Ren, L. Zhang, et al. 2016. A high density genetic map and QTL for agronomic and yield traits in Foxtail millet [Setaria italica (L.) P. Beauv.]. BMC Genomics 17: 336.
Guo, Y. W., Y. Q. Wu, J. A. Anderson, J. Q. Moss, L. Zhu, and J. M. Fu. 2017. SSR marker development, linkage mapping, and QTL analysis for establishment rate in common bermudagrass. Plant Genome 10: https://doi.org/10.3835/ plantgenome2016.07.0074.
Haupt, W., T. C. Fischer, S. Winderl, P. Fransz, and R. A. Torres-Ruiz. 2001. The CENTROMERE1 (CEN1) region of Arabidopsis thaliana: Architecture and functional impact of centromatin. Plant Journal 27: 285–296.
Hodel, R. D. G. J., M. C. Segovia-Salcedo, J. B. Landis, A. A. Crowl, M. Sun, X. X. Liu, M. A. Gitzendanner, et al. 2016. The report of my death was an exaggeration: A review for researchers using microsatellites in the 21st century. Applications in Plant Sciences 4(6): 1600025.
Hosouchi, T., N. Kumezawa, H. Tsuruoka, and H. Kotani. 2002. Physical map-based sizes of the centromeric regions of Arabidopsis thaliana chromosomes 1, 2, and 3. DNA Research 9: 117–121.
Hoşseinzadeh-Colagari, A., M. I. Haghhighatnia, Z. Amiri, M. Mohadjerani, and M. Tafiri. 2016. Microsatellite (SSR) amplification by PCR usually led to polymorphic bands: Evidence which shows replication slippage occurs in tend or nascent DNA strands. Molecular Biology Research Communications 5: 167–174.
Huson, D. H. 1998. SplitsTree: Analyzing and visualizing evolutionary data. Bioinformatics 14: 68–73.
Jarvis, D. E., Y. S. Ho, D. J. Lightfoot, S. M. Schmöckel, B. Li, T. J. A. Borm, H. Ohyanagi, et al. 2017. The genome of Chenopodium quinoa. Nature 542: 307–312.
Kim, D., J. M. Paggi, C. Park, C. Bennett, and S. L. Salzberg. 2019. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nature Biotechnology 37: 907–915.
Kloeper, T. H., and D. H. Huson. 2008. Drawing explicit phylogenetic networks and their integration into SplitsTree. BMC Evolutionary Biology 8: 22.
Lamesch, P., T. Z. Berardini, D. H. Li, D. Swarbreck, C. Wilks, R. Saadaharan, R. Muller, et al. 2012. The Arabidopsis Information Resource (TAIR): Improved gene annotation and new tools. Nucleic Acids Research 40: D1202–D1210.
Langella, O. 1999. Populations 1.2.31, CNRS UPR9034. 2002. Available from: http://www.pasteur.fr/environnement/laboratoire/populations/Populations.html (accessed 27 October 2020).
Li, H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:1303.3997 [preprint] [published 16 March 2013]. Available from https://arxiv.org/abs/1303.3997 [accessed 27 October 2020].
Li, H. 2018. Minimap2: Pairwise alignment for nucleotide sequences. Bioinformatics 34: 3094–3100.
Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, et al. 2009. The Sequence alignment/map (SAM) format and SAMtools. Bioinformatics 25(16): 2078–2079.
Appendix 1. Protocol for making a comprehensive SSR reference Fasta file using MISA, BEDTools, and several shell commands.

1. Run MISA:
   perl misa.pl my_Reference.fasta
   MISA requires a misa.ini file in the directory where MISA is being executed that should look like this:
   \n   definition(unit_size.min_repeats): 2-6 3-4 4-4
   interruptions(max_difference_between_2_SSRs): 100
   GFF: true

2. Modify the MISA-produced .gff files as follows:
   A. Remove any compound SSRs and calculate how much flanking sequence is available at each SSR locus:
   for i in *.gff; do awk '{if ($5-$4 >10 && $5-$4 <50) print $1 "\t" $4-100 "\t" $5+100}' > $1.mod.gff; echo "processing $i"; done
   *We note that the size of the flanking region can be changed using the awk statement above (simply change the "-100" and "+100" to the flanking size wanted).
   B. Concatenate the modified .gff files:
   for i in *.mod.gff; do cat $i >> cat_filter1.gff; done
   C. Remove SSRs that do not have sufficient flanking sequence:
   awk '($2 > 0)\{cat $1\}' > cat_filter1.gff
   D. Use BEDTools getfasta to make the SrSReferenceFile.fasta:
   bedtools getfasta -si my_Reference.fasta -bed cat_filter1.gff -fo SrSReferenceFile.fasta

Appendix 2. Protocol for BWA-MEM read mapping, PCR duplicate read removal, and read mapping quality control using SAMTools (Li et al., 2009).

1. Index the SrsReferenceFile.fasta:
   bwa index SrSReferenceFile.fasta

2. Map the Illumina reads to the SrsReferenceFile.fasta (paired-end reads process shown):
   for forward_file in _1.fq.gz; do name=echo $forward_file | sed 's/_1/1P/'; bwa mem -M -f SrSReferenceFile.fasta $name _1P.fq.gz $name _2P.fq.gz $name samName; done
   Note: Raw reads should be pre-trimmed. Using Trimmmomatic (Bolger et al., 2014) will produce a _1P.fq.gz and _2P.fq.gz file for each sample.

3. Remove PCR duplicate reads:
   A. Sort SAM files by read name for i in *.sam; do samtools sort -n -o $i.sorted $i; done
   B. Identify mate coordinates for i in *.sorted; do samtools fixmate -m $1 $1.fixmate; done
   C. Re-sort SAM files by genomic location for i in *.fixmate; do samtools sort -o $i.position $i; done
   D. Mark and remove duplicates:
   for i in *.position; do samtools markdup -r $i $i.markdup; done
   *Each sample will have a markdup file where PCR duplicates have been removed (the other files can be deleted).

4. Read mapping quality control:
   A. While the whole SAM file can be passed to SSRgenotyper, we encourage users to first filter the markdup file with SAMTools to improve performance and to remove errantly mapped reads. This command will remove reads with mapping quality < 45.
   For i in *.markdup; do SAMTools view $i -q 45 > $i.Q45.sam; done
   Note: SSRgenotyper provides further filtering (option -Q) that can be used for additional filter stringency if required.

5. A .txt file listing all of the quality-controlled SAM files (SamFiles.txt) to be processed by SSRgenotyper can be produced with:
   ls *.Q45.sam > samFiles.txt

Appendix 3. Commands, BioProject numbers, and SRA identifier utilized for each of the test data sets.

1. Run command for accuracy and error profile analysis, where -S was provided at various read coverage depths, ranging from 1 to 10 (C: quinoa: PRJNA4640095, SRA no.: SRR12041554–SRR12041555; A: atlantica: PRJNA546592):
   python3 SSRgenotyperV3.py SrSReferenceFile.fasta samFiles.txt quinoa -M 0.25 -R 4 -P 3 -B 3 -S -F 0.35 -f 0.30 -Q 60 -s 0.1 -m 0 -N 11 -L 0.25

2. Run command for the Arabopsis thaliana recombinant inbred line (RIL) population (BioProject no.: PRJNA418075, SRA no.: SRR6292944–SRR6293128; Serin et al., 2017):
   python3 SSRgenotyperV3.py SrSReferenceFile.fasta samFiles.txt BayreuthXShahdara -M 0.51 -R 4 -P 3 -B 3 -S -F 0.35 -f 0.30 -Q 60 -s 0.1 -m 0 -N 11 -L 0.25

3. Run command for the foxtail millet RIL population (BioProject no.: PRJNA562988, SRA no.: SRR10038747–SRR10038795; Liu et al., 2020):
   python3 SSRgenotyperV3.py SrSReferenceFile.fasta samFiles.txt foxtail -M 0.51 -R 4 -P 3 -B 3 -S -F 0.35 -f 0.30 -Q 60 -s 0.1 -m 0 -N 11 -L 0.25

4. Run command for the quinoa diversity panel (BioProject no.: PRJNA306026, SRA no.: SRR4300210–SRR4300229; Jarvis et al., 2017):
   python3 SSRgenotyperV3.py SrSReferenceFile.fasta samFiles.txt quinoa -M 0.25 -R 4 -P 3 -B 3 -S -F 0.35 -f 0.30 -Q 60 -s 0.1 -m 0 -N 11 -G