The Report of My Death was an Exaggeration: A Review for Researchers Using Microsatellites in the 21st Century

Authors: Hodel, Richard G. J., Segovia-Salcedo, M. Claudia, Landis, Jacob B., Crowl, Andrew A., Sun, Miao, et al.

Source: Applications in Plant Sciences, 4(6)

Published By: Botanical Society of America

URL: https://doi.org/10.3732/apps.1600025

Applications in Plant Sciences

Applications in Plant Sciences 2016 4(6): 1600025

Review Article

The report of my death was an exaggeration: A review for researchers using microsatellites in the 21st century[1]

Richard G. J. Hodel[2,3,7], M. Claudia Segovia-Salcedo[4], Jacob B. Landis[2,3], Andrew A. Crowl[2,3], Miao Sun[3], Xiaoxian Liu[2,3], Matthew A. Gitzendanner[2], Norman A. Douglas[2], Charlotte C. Germain-Aubrey[3], Shichao Chen[5], Douglas E. Soltis[2,3,6], and Pamela S. Soltis[3,6]

[2]Department of Biology, University of Florida, Gainesville, Florida 32611 USA; [3]Florida Museum of Natural History, University of Florida, Gainesville, Florida 32611 USA; [4]Departamento de Ciencias de la Vida, Universidad de las Fuerzas Armadas–ESPE, Sangolqui, Ecuador; [5]College of Life Sciences and Technology, Tongji University, Shanghai 200092, China; and [6]The Genetics Institute, University of Florida, Gainesville, Florida 32611 USA

Microsatellites, or simple sequence repeats (SSRs), have long played a major role in genetic studies due to their typically high polymorphism. They have diverse applications, including genome mapping, forensics, ascertaining parentage, population and conservation genetics, identification of the parentage of polyploids, and phylogeography. We compare SSRs and newer methods, such as genotyping by sequencing (GBS) and restriction site associated DNA sequencing (RAD-Seq), and offer recommendations for researchers considering which genetic markers to use. We also review the variety of techniques currently used for identifying microsatellite loci and developing primers, with a particular focus on those that make use of next-generation sequencing (NGS). Additionally, we review software for microsatellite development and report on an experiment to assess the utility of currently available software for SSR development. Finally, we discuss the future of microsatellites and make recommendations for researchers preparing to use microsatellites. We argue that microsatellites still have an important place in the genomic age as they remain effective and cost-efficient markers.

Key words: genotyping by sequencing (GBS); microsatellite development; next-generation sequencing (NGS); restriction site associated DNA sequencing (RAD-Seq); simple sequence repeats (SSR); transcriptomes.

Microsatellites, or simple sequence repeats (SSRs), are short repeated DNA motifs (typically one to six nucleotides) located throughout eukaryotic genomes (Li et al., 2002; Zane et al., 2002). Within microsatellite regions, these motifs are repeated several to dozens of times, although the number of repeats is highly variable (Selkoe and Toonen, 2006). Replication slippage is generally considered the mechanism that creates variation in the number of repeats (Ellegren, 2004). Microsatellites exhibit high levels of polymorphism and have a high mutation rate—between 10^{-3} and 10^{-4} per locus per generation, compared to approximately 10^{-9} nucleotides per generation for nucleotide substitutions across the entire genome in eukaryotes (Li et al., 2002). The high level of polymorphism in microsatellites makes these markers powerful tools for assessing genetic similarity between individuals or closely related taxa (Guichoux et al., 2011; Kalia et al., 2011). Since developing microsatellite loci (see Appendix 1 for a glossary of terms used in this paper) became cost-effective in the late 1990s, researchers have used them frequently in studies requiring high levels of polymorphism, generating approximately 225,000 published articles (search of Web of Science performed April 2016, term: microsatellite* OR "simple sequence repeat*").

Microsatellites have been used for a wide variety of applications, including genome mapping, forensics, parentage analysis, conservation genetics, identification of the parentage of polyploids, phylogeography, and population genetics (Ellegren, 2000; Esselink et al., 2004; Kalia et al., 2011). Their abundance in the genome, high levels of polymorphism, and cost effectiveness have contributed to the attractiveness of these markers. They are inexpensive when compared to the cost of using next-generation sequencing (NGS) techniques to generate sufficient data to differentiate among closely related individuals (Davey et al., 2011). Additionally, unlike with NGS data, the relatively small number of loci used in an SSR study means that each locus can be manually genotyped, reducing errors. Because they are PCR-based markers, microsatellite loci can be successfully amplified from poor-quality or low quantities of DNA, making them useful markers for studies involving ancient DNA or museum specimens (Wandeler et al., 2007). Many microsatellite primers will work in species closely related to the one for

[1]Manuscript received 4 March 2016; revision accepted 25 May 2016.

The authors thank three anonymous reviewers and *APPS* associate editor Dr. Mitch Cruzan for many helpful comments on previous versions of this manuscript, and Mark Twain for lending us part of our title. This work was supported in part by a National Science Foundation Doctoral Dissertation Improvement Grant (DEB-1501600 to D.E.S. and R.G.J.H.).

[7]Author for correspondence: hodel@ufl.edu

doi:10.3732/apps.1600025

which they were originally designed, allowing for multispecies studies.

Many of the applications noted above select microsatellites for their presumably neutral nature. SSRs can also be used in studies favoring nonneutrality; the association of microsatellites with a gene under selection can be used for the construction of genetic maps (Serikawa et al., 1992; Echt et al., 2011). Microsatellites are used in crop science and forestry to build high-density genetic maps useful for locating resistance to a pest or disease, or control for a desired trait (e.g., Hardwood Genomics Project; www.hardwoodgenomics.org). For example, a map with 19 microsatellites was built around *Ppr1*, a locus controlling *Puccinia psidii* rust resistance in *Eucalyptus* L'Hér. (Mamani et al., 2010). Also, SSRs were used to map blight resistance genes in *Castanea dentata* (Marshall) Borkh., the American chestnut (Jacobs et al., 2013). Finally, in the case of very large genomes, microsatellites are the favored method to construct a genetic map in the absence of a reference genome. The efforts to build genetic maps for gymnosperms have been successful with the association of single-nucleotide polymorphisms (SNPs) with SSRs in *Pinus taeda* L. (Echt et al., 2011).

The use of microsatellites, however, is not without concerns and caveats. The mechanism that leads to mutations in microsatellites (replication slippage) is prone to back mutations, promoting homoplasy (Viard et al., 1998). Extensive homoplasy leads to erroneous inferences of homology. Although the high potential for homoplasy can be modeled (e.g., using the stepwise mutation model), homoplasy complicates analyses and lowers confidence in inferences made using microsatellites (Slatkin, 1995). Furthermore, the high rates of polymorphism and homoplasy make microsatellites unsuitable for phylogenetic analyses beyond very closely related species (e.g., Soltis et al., 1998). Another concern is that the large number of alleles per locus associated with microsatellites can inflate *F*-statistic estimates relative to biallelic markers, such as SNPs (Whitlock, 2011). Conversely, in some cases, allele frequencies can also suppress *F*-statistic estimates in microsatellites: estimates of genetic diversity among populations (F_{ST}) are very low when the frequency of the most common allele is either very low or very high (Jakobsson et al., 2013). Additionally, genotyping errors, which can bias downstream analyses (Hoffman and Amos, 2005), are also potential concerns. Although NGS techniques such as genotyping by sequencing (GBS) and restriction site associated DNA sequencing (RAD-Seq) (Appendix 1) also have the potential for sequencing errors, the large amount of data generated with NGS methods diminishes this concern—effectively "drowning out" erroneous signal (Hou et al., 2015). Conversely, the relatively small number of loci used in traditional microsatellite studies means that genotyping errors can have a large downstream effect. The genomic age has ushered in a variety of new techniques that offer alternatives to SSRs. Thus, in this review of microsatellites, we address the following sets of questions:

1. How do SSR markers compare to NGS markers generated using GBS/RAD-Seq? What factors should researchers consider when choosing a genotyping method?
2. For researchers planning to use microsatellites, what details are critical when designing a project? What is the current state of SSR marker development?
3. What is the future of microsatellite markers? How should researchers use microsatellites in 2016 and beyond?

We first compare the advantages and disadvantages of using microsatellites as opposed to GBS/RAD-Seq. We then review techniques currently used for identifying microsatellite loci and developing primers, emphasizing those that make use of NGS approaches. Additionally, we make recommendations for researchers considering using microsatellites and address the question: Are SSRs a viable option when NGS techniques are rapidly becoming more cost-effective? We also review software packages for analyzing microsatellite data and make recommendations for researchers planning to use microsatellites.

MICROSATELLITES VS. GBS/RAD-SEQ

For a plant population geneticist beginning a study, there are important decisions to make regarding marker choice before collecting a single sample. Microsatellites have been, and still remain, a viable option for collecting genetic data, whereas GBS/RAD-Seq methods are increasing in popularity (Narum et al., 2013). Researchers need to consider carefully a variety of factors before beginning a study, including the project budget, the size of the group to be investigated (number of samples), the genetic resolution required, and the availability of genomic resources for the study group (e.g., a sequenced genome or other existing resources). When there is a very limited budget or only a small number of individuals can be included (e.g., a conservation genetic study on a rare species), microsatellites remain a good choice (Gardner et al., 2011). However, it may be preferable to start with GBS or RAD-Seq when beginning a long-term project, although samples must be organized into discrete groups for multiplexing, as the use of multiplexing is what makes these techniques affordable. Importantly, if additional data are needed, from the sequencing perspective, it would be as expensive to add one more sample as it would to add 100. Due to lane effects and other stochasticities associated with NGS, it is advisable to use standards in a long-term project that will use different sequencing machines. A strong background in computing skills and bioinformatics is needed to deal with the large quantity of data generated by NGS approaches, whereas researchers can complete microsatellite analysis with limited computing skills and/or resources on a laptop computer using one or more graphical user interface (GUI) programs.

RAD-Seq and GBS are approaches that combine the value of reducing genome complexity with restriction enzymes (REs) and NGS-based SNP discovery and genotyping (Davey and Blaxter, 2010; Davey et al., 2011; Etter et al., 2011; Arnold et al., 2013; Andrews et al., 2016). These methods enable discovery of thousands of markers, even in nonmodel organisms, and characterization of different levels of genetic variation across the genome (Hohenlohe et al., 2010; Rowe et al., 2011; Liu et al., 2013; Lu et al., 2013). The main differences between RAD-Seq and GBS are methodological, relating to which REs are used to digest DNA, how sequencing adapters and multiplexing barcodes are added to samples, and the use of a size selection step (Elshire et al., 2011; Cronn et al., 2012). Hereafter, we will treat RAD-Seq and GBS as a suite of methods united by their use of REs to reduce genome complexity prior to multiplexed NGS and will refer to this suite of methods as RAD/GBS. Library complexity is directly related to genome complexity and size and the choice of REs (Beissinger et al., 2013). With RAD/GBS, there is a trade-off between the number of SNPs and coverage of each locus, which can be mediated by choosing REs with longer recognition sites, resulting in higher coverage of

fewer loci. This approach enables the use of these data for population genetics (Beissinger et al., 2013; Lu et al., 2013; Narum et al., 2013).

The primary advantage of RAD/GBS is that thousands of loci can be simultaneously generated for hundreds of individuals, with costs as low as US$35 per sample (assuming strategic sharing of REs, adapters, barcodes, and efficient multiplexing with an optimal number of samples). Reducing genome complexity with REs is a very specific, fast, and simple procedure (Sonah et al., 2013; Andrews et al., 2016). There is no requirement for a priori knowledge of the genome of the species; however, a reference genome facilitates selecting an appropriate RE (Sonah et al., 2013; Spindel et al., 2013; Liu et al., 2014). REs can be chosen that prevent highly repetitive regions and target low-copy regions, increasing the efficiency of the research goals and reducing computational time with alignment procedures. A multiplex barcoding system increases efficiency and reduces costs (Smith et al., 2010; Andolfatto et al., 2011; Elshire et al., 2011; Sonah et al., 2013). SNPs also have a number of advantages when directly compared to SSRs: they are less prone to homoplasy than SSRs and are also easier to locate in most single-copy regions of the genome than SSRs (Rafalski, 2002). Another advantage of SNPs is that relatively few SNPs are needed to define a haplotype or to detect linkage disequilibrium (Rafalski, 2002).

RAD/GBS approaches have several disadvantages as well. Problems may result from: (1) the frequent conflation of paralogous loci due to misassembly of reads (Etter et al., 2011; Xu et al., 2014), (2) sequencing errors and inaccurate genotyping with low sequencing depths (Arnold et al., 2013), (3) PCR bias in library construction (Arnold et al., 2013), and (4) nonrandom cleavage by enzyme digestion (Arnold et al., 2013). The first three issues have largely been addressed by improvements in algorithms and software for processing loci, improvements in sequencing technology and careful multiplexing, and multiple PCR steps, respectively. However, sampling DNA based on REs may still include a bias in allele frequency estimation. Mutations in restriction sites can lead to underestimating diversity and introduce genealogical biases, causing haplotypes to be nonrandomly sampled (Arnold et al., 2013). Additionally, the nucleotide composition of the restriction site affects which areas of the genome are sampled; the goals of the study should guide which REs are chosen. GC content should be carefully considered when selecting REs, as GC-rich REs lead to overrepresentation of the portions of the genome high in GC content (DaCosta and Sorenson, 2014). Additionally, RAD/GBS data often overestimate heterozygosity (Arnold et al., 2013; Gautier et al., 2013). Unlike with SSR markers, manual validation is impractical with RAD/GBS data, and biases or errors may be impossible to detect (Etter et al., 2011; Davey et al., 2013).

Several aspects of RAD/GBS present challenges that researchers need to consider. Multiplex sequencing protocols for RAD/GBS often depend on an accurate quantification of high-molecular-weight DNA (Elshire et al., 2011). However, this requirement may be waning, as recent studies have used RAD/GBS on herbarium specimens, which may have degraded DNA (Beck and Semple, 2015). Little information is currently available about how markers discovered with RAD/GBS are distributed across the genome, although studies in wheat and barley suggest that these markers are uniformly spaced (Poland et al., 2012). Large variation in GC content among taxa may introduce biases, leaving important genomic regions over- or underrepresented (Beissinger et al., 2013). However, large differences in GC content among close relatives are unusual, meaning this will likely not be an issue in population genomic studies. Another concern is that RAD/GBS data sets often have a huge amount of missing data compared to traditional genotyping methods. Researchers must make critical decisions about whether to exclude loci and/or individuals from analyses when there are high levels of missing data. Another consideration is that missing data are not randomly spread across individuals and/or loci due to the nature of the genomic library construction (REs). Therefore, allelic dropout, geneological biases, and underestimation of diversity may be some of the consequences of missing data in RAD/GBS methods. Aspects of library construction, data processing, and the divergence history of study species may affect results; simulations and more studies are needed to define guidelines about how to handle missing data when using RAD/GBS (Huang and Knowles, 2016).

Whereas RAD/GBS are powerful methods for diploid species, many challenges remain for calling SNPs in polyploids. Specialized SNP genotyping algorithms are required when using RAD/GBS in polyploids (Narum et al., 2013). Because sequencing coverage determines the level of missing data, the large genomes of some plants, especially polyploids, can lead to low coverage. In all RAD/GBS protocols, the average number of reads per sample will be based on multiplexing and the number of independent sequences generated by the sequencing platform—either sequencing coverage or number of samples multiplexed will be reduced in polyploids as compared to diploids (Poland and Rife, 2012). Large plant genomes, due to either repetitive DNA or polyploidy, can lead to the erroneous construction of artifactual composite "loci" with falsely inferred polymorphisms. Longer reads facilitate the discovery of more polymorphisms when RAD/GBS is applied to polyploids, which require genome-specific polymorphisms to differentiate among homeologous sequences (Poland and Rife, 2012; Sonah et al., 2013).

One of the most important criteria for selecting a method is cost-feasibility; we present two approximate budgets (Appendix 2) for genotyping 96 individuals: one that involves developing and genotyping microsatellites and one that implements RAD/GBS. As of May 2016, if a researcher needs to develop his/her own microsatellite loci, the cost of genotyping approximately 96 individuals using 12–15 microsatellite markers is similar to performing RAD/GBS on 96 individuals. It is very challenging to present a budget that accounts for all the factors that will determine the cost of a project, but we attempt some approximate budgets that can be used as guidelines when designing projects.

MICROSATELLITE DEVELOPMENT: REVIEW OF TECHNIQUES

If microsatellite markers are the chosen approach, researchers have two options: generate sequence data for microsatellite detection or mine pre-existing resources for marker discovery. The first option requires decisions on library preparation, sequencing platform (including read length and depth), and software for marker detection. The second option makes the first two decisions unnecessary and bypasses sequencing costs, but software choice is still important.

Historical methods of microsatellite library construction—Microsatellite libraries were traditionally developed by digestion with one or more REs (Ritschel et al., 2004). A linker of known sequence would be ligated onto the digested fragments,

and one or more probes containing repeat sequences were hybridized to those fragments. This enrichment step limited the nature of the microsatellites that would ultimately be obtained at the end of the procedure. The repeat-enriched fragments were then recovered using streptavidin-coated beads (Nunome et al., 2006). The library was amplified and the PCR products cloned and sequenced. The enrichment strategy is time-consuming (10–14 d), and the DNA extracted for such a protocol has to be of high quality and quantity. The yield of such a library construction is typically eight to 20 polymorphic loci for 30–60 SSR primer pairs tested (Zalapa et al., 2012), and the initial cost is low (less than US$500 for a cloning kit).

NGS has transformed the development of microsatellite loci for ecological and evolutionary studies. Current approaches allow quick and inexpensive identification of large numbers of loci in nonmodel organisms. Studies so far have largely focused microsatellite discovery efforts on the Roche 454 (454 Life Sciences, a Roche Company, Branford, Connecticut, USA) and Illumina (Illumina, San Diego, California) platforms (Jennings et al., 2011; Zalapa et al., 2012), although Pacific Biosciences (PacBio, Menlo Park, California, USA) (Grohme et al., 2013; Wei et al., 2014) and Ion Torrent (Thermo Fisher Scientific, Waltham, Massachusetts, USA) (Huey et al., 2013; Kameyama and Hirao, 2014) have also been used. Because read length greatly affects the ability to discover microsatellite markers, as longer reads will more likely include the flanking regions needed for primer design (Lepais and Bacles, 2011; Schoebel et al., 2013; Elliott et al., 2014), the 454 sequencing platform was used extensively for microsatellite development (Castoe et al., 2010). On a per-megabase basis, however, 454 is less cost-effective than Illumina (Glenn, 2011; Appendix S1). Between January 2013 and April 2016, 74 projects using 454 were published in *Applications in Plant Sciences*, yielding between eight and 91 polymorphic loci, with an average of 16 loci, derived from an average of 139,418 reads. Roche announced they will be discontinuing the use of the 454 instrument in 2016. Future projects using NGS to develop microsatellite loci will rely on alternative platforms.

***Current library preparation methods*—**Several approaches can reduce genomic complexity and enrich for microsatellites prior to library building (Glenn, 2011). Method selection depends on platform throughput, number of individuals, desired coverage, and availability of a reference genome or transcriptome (Jennings et al., 2011). Microsatellite-enrichment methods require a priori decisions on the type of repeat motif and size of repeat sequence, creating bias in locus choice (Castoe et al., 2010). Using shotgun sequencing to identify loci allows for random sampling of the genome and is preferable to microsatellite-enrichment techniques. Regardless of sequencing platform and library preparation, however, NGS approaches to microsatellite discovery are more time- and cost-effective and provide more potential loci than traditional approaches. The limiting step for microsatellite studies is no longer marker discovery and development, but instead, screening and validation of loci (Wei et al., 2014).

The short read lengths obtained with platforms such as Illumina and Ion Torrent previously limited their utility for microsatellite development. However, as Illumina platforms generate longer read lengths (MiSeq currently generates 2 × 300 bp reads), this limitation is changing. Zalapa et al. (2012) reported two of 17 projects in their analysis used Illumina platforms. Between January 2013 and April 2016, 28.8% (34 of 118) of primer notes published in *Applications in Plant Sciences* utilizing NGS used Illumina. For studies using Illumina, the average number of polymorphic microsatellite markers reported was 15 loci, and the average number of potential loci per study was 15,539, which is larger than other platforms (e.g., 454, with an average of 4400 potential markers). This is predominantly due to the greater throughput of Illumina (see Appendix S1).

***Sequencing platform*—**Read length, read output, and error rate all affect platform choice for generating sequence data for marker discovery (Glenn, 2014; Appendix S1). Currently there are three Illumina platforms available: MiSeq, HiSeq, and NextSeq, with the HiSeq ×10 debuting in 2016. The MiSeq, which only has a single lane, has the fastest run times and the longest read lengths (~56 h for 2 × 300 bp). However, the MiSeq output consists of relatively few reads (50 million) of up to 2 × 300 bp at a higher cost per mega base pair compared to the HiSeq. The HiSeq has a low cost per megabase of data—up to 500 gigabytes (GB) of data per flow cell. However, these reads are shorter than the MiSeq; until recently, the longest was 2 × 150 bp, and the runs take up to six days; however, the new HiSeq v2 reagents allow 2 × 250 bp in rapid run mode. Drawbacks to the HiSeq are the requirement to fill all eight lanes before running, and that a single flow cell can be processed only as a rapid run or a high-throughput run. The NextSeq falls between the two other platforms in performance; it can generate reads of 2 × 150 bp, with a high-throughput run generating up to 120 GB of data in ~29 h. All three models have a low final error rate of 0.1% (primarily substitution-type miscalls; Glenn, 2014).

Two additional platforms that are increasing in use for SSR discovery are Ion Torrent and PacBio. Ion Torrent has three chip options generating between 50 Mbp and 2 Gbp of data, with read lengths of 200 or 400 bp, and sequencing time ranging between 3 and 7.9 h. The PacBio platform is a single-molecule real-time sequencer, which removes PCR errors that can be introduced when using other platforms. Of the three platforms reviewed, PacBio has the greatest flexibility in run times (30 min to 6 h per single-molecule real-time sequencing [SMRT] cell) and run size (one to 16 SMRT cells) and provides the longest read lengths, up to 20 Kb—an attractive feature for microsatellite discovery. PacBio suffers from the highest error rate—approximately 13% in raw reads. However, unlike Illumina and Ion Torrent, these errors are stochastic, meaning that a final error rate of less than 1% can be achieved in the consensus sequence of numerous raw reads (Glenn, 2014). Unfortunately, the advantages of the PacBio system come at a cost—it delivers a very low total number of reads per run (500 Mbp to 1 Gbp per SMRT cell) and a high cost per Mbp of data (Appendix S1).

Many pipelines have been published using paired-end Illumina reads (e.g., Miller et al., 2013; Andersen and Mills, 2014), with genomic DNA or RNA-Seq data. Gilmore et al. (2013) estimated the time and cost of using Illumina data to produce markers from eight samples to be approximately 20 h of laboratory work for sample preparation and approximately US$51 per sample. Several recent studies have justified the use of other platforms, mainly Ion Torrent (Elliott et al., 2014) and PacBio (Wei et al., 2014). In a comparison of the utility of 454 and Ion Torrent, Elliott et al. (2014) found the Ion Torrent recovered shorter microsatellite repeats (due to shorter reads), but more markers were discovered at a lower cost and more quickly than with 454. The PacBio RS platform may become a preferred method for obtaining highly variable SSRs in the future, especially if error rates and price decrease; the latter is proposed with

their Sequel instrument. Small-scale marker development results in long reads using a single SMRT cell, which may yield thousands of repeats (Grohme et al., 2013; Wainwright et al., 2013).

Mining existing data sets—Another option for developing microsatellite markers is using publicly available sequence data from online repositories such as the Sequence Read Archive (SRA; http://www.ncbi.nlm.nih.gov/sra). This archive houses a large collection of raw sequence data from various NGS platforms and approaches such as targeted-gene capture, genome skimming, restriction digests, and transcriptome sequencing. To determine the potential of these data sets to generate microsatellite markers, targeted-gene capture (SRR2658270; Landis et al., 2015) and RAD-Seq reads (Hodel et al., unpublished data) were used for marker discovery. In both data sets, over 100,000 potential loci were discovered, highlighting the utility of publicly available data for mining SSR loci (Tables 1 and 2). Another resource for researchers is the One Thousand Plant Transcriptomes Project (1KP; www.onekp.com; Matasci et al., 2014), which has transcriptome assemblies for over 1000 plant species. The companion paper to our review presents over five million SSR loci that can be used in thousands of plant species (Hodel et al., 2016). It is important to note that once potential loci are identified from NGS data, this is just the starting point for developing a functional microsatellite genotyping system and extensive and costly screening of loci will be required, as outlined in the budget in Appendix 2.

SOFTWARE FOR MICROSATELLITE DEVELOPMENT

Once researchers generate or obtain NGS data, the next step is to use a software program to identify potential loci to screen. We tested the effectiveness and ease of use of 10 commonly implemented software packages for microsatellite identification using four *Arabidopsis thaliana* NGS data sets mined from SRA. The data sets are: a single-end (1 × 100 bp) lane of Illumina HiSeq 2000 (ERR368422), which is 10.9 million reads and a total of 1.5 Gbp of sequence, a paired-end (2 × 100 bp) Illumina HiSeq 2000 lane (ERR965681; 97 million reads and a total of 8.7 Gbp of sequence), a paired-end (2 × 250 bp) Illumina MiSeq run (ERR365834; 13.2 million reads and a total of 3.3 Gbp of sequence), and a PacBio sequencing run (SRR1284764; 476 Mbp of sequence in 163,500 reads). We obtained the data in FASTA and FASTQ files from SRA using the SRA toolkit. Hereafter, these data sets will be referred to as HiSeq1, HiSeq2, MiSeq, and PacBio. FASTA files for each data set ranged in size from 445 MB (PacBio) to 5.7 GB (HiSeq2). For some software packages, we had to use other file formats (e.g., FASTQ), but we report FASTA file sizes for simplicity.

We selected these four data sets to investigate how read number, read length, sequencing platform, and data set size affected the performance of each software package. Our goal was to provide readers with the information necessary to obtain microsatellite loci from publicly available data as easily as possible. We ran each data set through each software program, using the same settings in each program as much as possible. We selected the default values from QDD3 to use in every program, as the default values were difficult to change in QDD3. Although it is important to use a consistent set of parameters for every program, the actual parameters used can be arbitrary, so we used QDD3 defaults. The critical parameters to standardize were the number of repeats of a certain length motif required to call a locus. The QDD3 default values are: homopolymers, 1,000,000 repeats; dinucleotides, five repeats; trinucleotides, five repeats; tetranucleotides, five repeats; pentanucleotides, five repeats; hexanucleotides, five repeats. For each software package that ran to completion for all data sets, we report the total number of SSR loci found, the number of loci per mega base pair of sequence, and the distribution of loci across size motifs (di-, tri-, tetra-, penta-, hexanucleotides).

First, we summarize the utility and main characteristics of the software packages (see below, and Table 3). Next, we compare software packages, so future researchers are well-equipped to develop SSR loci easily. The goal of most of these programs is to search for SSR loci, quantify the distribution of loci across size motifs, and facilitate primer design. Many of these software packages use a GUI, but some are command line only and require knowledge of Perl or Python for software installation and execution. Many of the software packages interact with Primer3 (Rozen and Skaletsky, 1999) for primer design. Most programs are open source, platform independent, and capable of handling genomic data. When possible, we ran these software packages on a high-performance computing cluster. As noted below, some software packages would not run after a reasonable period of effort by a biologist proficient in command line and at least one programming language. We briefly describe and evaluate each program, report the resources required to run each one, how long execution took, and other relevant details for evaluating software packages (Table 3).

Most of the tested software packages executed successfully for all four test data sets and produced results consistent with other programs (Tables 4 and 5). HighSRR and SSR_pipeline

TABLE 1. The number of loci found in an SSR search and the number of loci found per mega base pair sequence for each software package in each of two data sets used to highlight the vast potential resources available for researchers who cannot generate their own sequence data to search for SSRs.

	Rhizophora RAD-Seq[a]		*Saltugilia* gene capture[b]	
Software package	Total no. of loci	Loci/Mbp sequence	Total no. of loci	Loci/Mbp sequence
GMATo	448,569	39.3	181,223	671.2
MISA	448,746	39.4	181,449	672.0
MSATCOMMANDER	372,436	32.7	151,455	560.9
PAL_FINDER	NA	NA	140,463	520.2
Phobos (Geneious, STAMP)	450,948	39.6	181,616	672.7
SSR Locator	447,185	39.2	180,763	669.5

Note: NA = not applicable.
[a] 114 million single-end reads (1 × 100); 11.4-Gbp sequence; FASTA file: 12.5 GB.
[b] 900,000 paired-end reads (2 × 150; 1.8 million total reads); 270-Mbp sequence; FASTA file: 275 MB.

TABLE 2. The number and percentage of each repeat motif type using each software package found in the SSR search for each test data set.

Software package	*Rhizophora* RAD-Seq[a]	*Saltugilia* gene capture[b]
GMATo		
No. of dinucleotides (%)	335,835 (74.9)	103,688 (57.2)
No. of trinucleotides (%)	93,812 (20.9)	74,129 (40.9)
No. of tetranucleotides (%)	15,997 (3.6)	610 (0.3)
No. of pentanucleotides (%)	2233 (0.5)	307 (0.2)
No. of hexanucleotides (%)	692 (0.2)	2489 (1.4)
MISA		
No. of dinucleotides (%)	335,883 (74.8)	103,768 (57.2)
No. of trinucleotides (%)	93,933 (20.9)	74,208 (40.9)
No. of tetranucleotides (%)	16,005 (3.6)	612 (0.3)
No. of pentanucleotides (%)	2233 (0.5)	322 (0.2)
No. of hexanucleotides (%)	692 (0.2)	2539 (1.4)
MSATCOMMANDER		
No. of dinucleotides (%)	284,725 (76.4)	83,377 (55.1)
No. of trinucleotides (%)	74,305 (20.0)	65,374 (43.2)
No. of tetranucleotides (%)	11,409 (3.1)	473 (0.3)
No. of pentanucleotides (%)	1631 (0.4)	197 (0.1)
No. of hexanucleotides (%)	366 (0.1)	2034 (1.3)
PAL_FINDER		
No. of dinucleotides (%)	NA	83,421 (59.4)
No. of trinucleotides (%)	NA	54,787 (39.0)
No. of tetranucleotides (%)	NA	589 (0.4)
No. of pentanucleotides (%)	NA	313 (0.2)
No. of hexanucleotides (%)	NA	2053 (1.5)
Phobos (Geneious, STAMP)		
No. of dinucleotides (%)	336,595 (74.6)	103,807 (57.2)
No. of trinucleotides (%)	95,423 (21.2)	74,311 (40.9)
No. of tetranucleotides (%)	16,005 (3.5)	613 (0.3)
No. of pentanucleotides (%)	2233 (0.5)	322 (0.2)
No. of hexanucleotides (%)	692 (0.2)	2563 (1.4)
SSR Locator		
No. of dinucleotides (%)	334,836 (74.9)	103,599 (57.3)
No. of trinucleotides (%)	93,427 (20.9)	73,938 (40.9)
No. of tetranucleotides (%)	15,998 (3.6)	604 (0.3)
No. of pentanucleotides (%)	2232 (0.5)	290 (0.2)
No. of hexanucleotides (%)	692 (0.2)	2332 (1.3)

Note: NA = not applicable.

[a] 114 million single-end reads (1 × 100); 11.4-Gbp sequence; FASTA file: 12.5 GB.

[b] 900,000 paired-end reads (2 × 150; 1.8 million total reads); 270-Mbp sequence; FASTA file: 275 MB.

did not run to completion. The software packages that failed to run or complete the loci search were either old or not compatible with current NGS data sizes and formats. For instance, there are several types of FASTQ formats, but SSR_pipeline recognized only one old version, and HighSSR is unable to run with files larger than 2 GB. Other packages, including GMATO, PAL_FINDER, QDD3, SRR Locator, and STAMP, had limitations. These packages were either slow, could not handle all data types and/or sizes, or were difficult to use (e.g., they required a substantial amount of file formatting and manipulating). PAL_FINDER and MSATCOMMANDER (Faircloth, 2008) consistently found fewer loci than other software packages (Table 4). We recommend using Phobos (either by itself or through Geneious if Primer3 integration is desired) or MISA. We base these recommendations on ease of use and reliability of results.

Geneious is a desktop software suite for the organization and analysis of sequence data in molecular biology (Kearse et al., 2012). Microsatellite development requires several plugins (e.g., Phobos, Primer3, and MISA) to meet users' specific needs. It is commercial software, which requires purchasing a license for activation, raising the research budget. The component that searches for microsatellite loci is **Phobos**, which can be run independently of Geneious for free. Phobos has both GUI and command-line interfaces, and it processes large files quickly. Every data set tested completed the search in less than an hour on a standard laptop (2.5-GHz Intel Core i5, 8 GB RAM). Phobos does not interact directly with Primer3, but if Phobos is used through Geneious, the results of the loci search in Phobos can be easily piped to Primer3. For microsatellite loci development, Phobos is fast and user-friendly.

GMATo comes with a Java graphical interface and is ready to execute immediately after downloading (Wang et al., 2013). GMATo results are presented as a table of SSR loci statistics. It runs quickly; for the HiSeq2 data set (a 5.7-GB file), it completed the job within 52 min on a desktop Windows machine (eight Core 3.4-GHz Intel Core i7-2600 CPU, 16 GB RAM). However, the user cannot control the distribution of repeat number motifs—every repeat length must be set to the same value. This program is not capable of primer design, marker generation, or electronic mapping markers.

HighSSR detects microsatellites and eliminates redundancy in the PCR primers for recovered loci (Churbanov et al., 2012). It identifies and scores SSRs in raw sequencing reads with Tandem Repeats Finder (TRF; Benson, 1999) and stores them in a PostgreSQL database, reporting summary statistics, such as the number of alleles of each SSR locus, which can be analyzed by other software. HighSSR demultiplexes pooled libraries, assesses locus polymorphism, and implements Primer3 for primer design. Finally, MUSCLE (Edgar, 2004) is used to refine crude clusters and distill loci from them. However, it requires a Java virtual machine and access to a database on a PostgreSQL server. Moreover, nonuniversal parameter settings and various Java codes and shell scripts make it difficult to use. For the TRF executable file, we could only open our smallest test data file (PacBio; 445 MB).

MISA is short for MIcroSAtellite identification tool, which was originally designed to generate SSR loci from EST data (Thiel et al., 2003). It works immediately if Perl is installed and runs rapidly; the 5.7-GB HiSeq2 data set finished in 1.8 h (one node, one processor, and 4 GB of memory). Users are able to change the default settings by editing a configuration file (misa.ini), and MISA is able to generate primers. Its results are in tabular form, giving a summary of different statistics, such as the frequency of a specific microsatellite type. However, some studies indicate that MISA may have mined redundantly in overlapped microsatellites (e.g., Wang et al., 2013; Hodel et al., 2016).

MSATCOMMANDER enables rapid and automated microsatellite detection, locus-specific primer design, and tagging (Faircloth, 2008). It requires Python and writes output files in comma-separated value (CSV) format. However, the results are difficult to view and do not include general summary statistics about the types of microsatellite loci found. The user must spend considerable time filtering the output file to determine basic statistics (e.g., the number of dinucleotide repeats found). It utilizes Primer3 as its primer design and primer-tagging engine.

PAL_FINDER finds microsatellite repeat elements directly from raw NGS sequencing reads and then designs PCR primers to amplify these repeat loci (potentially amplifiable loci [PAL]) by interaction with Primer3 (Castoe et al., 2012). This is command-line software, which can be freely modified by the user via the required config file. However, its performance is very sensitive to data coverage (quantity and quality of PALs; Castoe

TABLE 3. Description of software packages used in this study, including operating systems, important features, URL where software can be obtained, number of citations, authors, and brief comments describing the ease of use.

Software	Operating system	Features	URL	Citations (Web of Science/Google Scholar)	Reference	Comments
Geneious	Linux, Mac OSX, Windows	Integrates multiple functions with plugins, user-friendly interface	http://www.geneious.com/features/microsatellite-analysis	395/633	Kearse et al., 2012	Very user friendly, but requires a paid license to run. The microsatellite development plugin (Phobos, http://www.ruhr-uni-bochum.de/ecoevo/cm/cm_phobos.htm) is freely available, very easy to use, and quick.
GMATo	Linux, Mac OSX, Windows	Both GUI and command line interface; SSR mining and statistics at genome level	http://sourceforge.net/projects/gmato/files/?source=navbar	0/8	Wang et al., 2013	Runs quickly and has clear output, but it is hard for the user to change important parameter settings.
HighSSR	Linux, Mac OSX, Windows	A Java program is designed for NGS data and capable of microsatellite detection, elimination of redundancy and primer development, and interacting with PostgreSQL, MUSCLE, and Primer3.	https://code.google.com/p/highssr/	7/12	Churbanov et al., 2012	Cannot open large files (>2 GB); unsuitable for most NGS data.
MISA	Linux, Mac OSX, Windows	Preprocessing sequences, motif search, and interacts with Primer3 for primer designs	http://pgrc.ipk-gatersleben.de/misa/	669/1150	Thiel et al., 2003	Fast, easy to configure, generates primers.
MSATCOMMANDER	Linux, Mac OSX, Windows	Motif search, interacts with Primer3 for primer design, and primer auto-tag	http://code.google.com/p/msatcommander/	428/509	Faircloth, 2008	Output is difficult to view; requires lots of filtering to find basic statistics.
PAL_FINDER	Linux, Mac OSX, Windows	Identifies and characterizes microsatellite repeat loci from shotgun genomic sampling by 454 or Illumina paired-end reads, and designs PCR primers by interacting with Primer3	http://sourceforge.net/projects/palfinder/ or http://www.snakegenomics.org/CastoeLab/Software.html	87/115	Castoe et al., 2012	Slow, struggles with large files; compatibility issues with many FASTQ formats; would not complete with the largest data set.
QDD3	Windows and Linux	A computer program to select microsatellite markers from raw sequence reads obtained from 454 or Illumina and design primers from large sequences at genomic level, dealing with the essential bioinformatics and equipped with both command line and a user-friendly graphical interface on the Galaxy server.	http://net.imbe.fr/~emeglecz/qdd	3/9	Meglécz et al., 2014	Relatively long running time; user cannot easily change parameter settings.
SSR Locator	Windows	Integrates SSR searches, frequency of occurrence of motifs and arrangements, primer design, PCR simulation, global alignments, and identity and homology searches; eliminates overlaps between adjacent sequences; interacts with Primer3	http://www.hindawi.com/journals/ijpg/2008/412696/	0/98	Da Maia et al., 2008	Requires tedious file reformatting; only available for Windows.
SSR_pipeline	Linux, Mac OSX, Windows	Identifies simple sequence repeats, e.g., microsatellites from paired-end high-throughput Illumina DNA sequencing data	http://pubs.usgs.gov/ds/778/	3/3	Miller et al., 2013	We could not get it to run after ~24 h of effort.
STAMP	Linux, Mac OSX, Windows	Based on STADEN package, with comprehensive integration of a set of extension modules to facilitate the processing of microsatellite markers, like Phobos, TROLL, Primer3, SQLite module.	http://www.awi.de/en/research/scientific_computing/bioinformatics/software/	11/15	Kraemer et al., 2009	Powerful, but not user friendly. Other modules associated with it (Phobos) quickly and conveniently locate potential loci.

TABLE 4. Software packages, the number of loci they find in an SSR search, and the number of loci they find per mega base pair sequence in each of the four test data sets for four sequencing platforms (MiSeq, HiSeq1, HiSeq2, PacBio).

	MiSeq (ERR365834)[a]		HiSeq1 (ERR368422)[b]		HiSeq2 (ERR965681)[c]		PacBio (SRR1284764)[d]	
Software package	Total no. of loci	Loci/Mbp sequence	Total no. of loci	Loci/Mbp sequence	Total no. of loci	Loci/Mbp sequence	Total no. of loci	Loci/Mbp sequence
GMATo	482,084	146.1	171,016	114.0	722,636	83.1	104,630	219.8
MISA	482,336	146.2	171,095	114.1	723,062	83.1	104,778	220.1
MSATCOMMANDER	388,663	117.8	135,168	90.1	543,610	62.5	82,588	173.5
PAL_FINDER	310,495	94.1	158,163	105.4	591,617	68.0	48,831	102.6
Phobos (Geneious, STAMP)	483,037	146.4	172,309	114.9	723,917	83.2	104,896	220.4
SSR Locator	481,863	146.0	170,934	114.0	722,580	83.1	104,120	218.7

[a] 6.6 million paired-end reads (2 × 250; 13.2 million total reads); 3.3-Gbp sequence; FASTA file: 3.9 GB.
[b] 10.9 million single-end reads (1 × 100); 1.5-Gbp sequence; FASTA file: 2.2 GB.
[c] 48.5 million paired-end reads (2 × 100; 97 million total reads); 8.7-Gbp sequence; FASTA file: 5.7 GB.
[d] 163,500 reads; 476-Mbp sequence; FASTA file: 445 MB.

et al., 2012). After approximately 24 h of effort manipulating FASTQ input files, we were unable to get the FASTQ mode to work. We could use any type of FASTA file in the "454" mode, including paired-end Illumina data, as long as all the reads were in a single file. This program has a slow run time relative to other software packages reviewed (>24 h for data sets >4 GB on a standard laptop [2.5-GHz Intel Core i5 with 8 GB RAM]).

QDD3 is composed of four separately running modules, with functions of quality trimming, microsatellite detection, redundancy removal, primer design, contamination checking, and

TABLE 5. The number and percentage of each repeat motif type found in the SSR search in each of the four test data sets for four sequencing platforms (MiSeq, HiSeq1, HiSeq2, PacBio).

Software package	MiSeq (ERR365834)[a]	HiSeq1 (ERR368422)[b]	HiSeq2 (ERR965681)[c]	PacBio (SRR1284764)[d]
GMATo				
No. of dinucleotides (%)	395,657 (82.1)	123,902 (72.5)	565,192 (78.2)	95,584 (91.4)
No. of trinucleotides (%)	82,874 (17.2)	42,764 (25.0)	151,596 (21.0)	8366 (8.0)
No. of tetranucleotides (%)	2333 (0.5)	2290 (1.3)	3390 (0.5)	556 (0.5)
No. of pentanucleotides (%)	525 (0.1)	803 (0.5)	895 (0.1)	99 (0.1)
No. of hexanucleotides (%)	695 (0.1)	1257 (0.7)	1563 (0.2)	25 (0.0)
MISA				
No. of dinucleotides (%)	395,740 (82.0)	123,918 (72.4)	565,328 (78.2)	95,634 (91.3)
No. of trinucleotides (%)	83,016 (17.2)	42,817 (25.0)	151,850 (21.0)	8454 (8.1)
No. of tetranucleotides (%)	2357 (0.5)	2294 (1.3)	3406 (0.5)	564 (0.5)
No. of pentanucleotides (%)	525 (0.1)	806 (0.5)	905 (0.1)	99 (0.1)
No. of hexanucleotides (%)	698 (0.1)	1260 (0.7)	1573 (0.2)	27 (0.0)
MSATCOMMANDER				
No. of dinucleotides (%)	325,676 (83.8)	99,465 (73.6)	432,335 (79.5)	77,096 (93.4)
No. of trinucleotides (%)	60,629 (15.6)	32,118 (23.8)	107,818 (19.8)	5148 (6.2)
No. of tetranucleotides (%)	1613 (0.4)	1824 (1.3)	1925 (0.4)	286 (0.3)
No. of pentanucleotides (%)	313 (0.1)	650 (0.5)	619 (0.1)	45 (0.1)
No. of hexanucleotides (%)	432 (0.1)	1111 (0.8)	913 (0.2)	13 (0.0)
PAL_FINDER				
No. of dinucleotides (%)	251,678 (81.1)	114,219 (72.2)	460,072 (77.8)	41,581 (85.2)
No. of trinucleotides (%)	56,389 (18.2)	40,088 (25.3)	126,509 (21.4)	6595 (13.5)
No. of tetranucleotides (%)	1570 (0.5)	2042 (1.3)	2909 (0.5)	531 (1.1)
No. of pentanucleotides (%)	359 (0.1)	717 (0.5)	774 (0.1)	98 (0.2)
No. of hexanucleotides (%)	499 (0.2)	1097 (0.7)	1353 (0.2)	26 (0.1)
Phobos (Geneious, STAMP)				
No. of dinucleotides (%)	396,367 (82.1)	124,755 (72.4)	566,081 (78.2)	95,743 (91.3)
No. of trinucleotides (%)	83,088 (17.2)	43,156 (25.0)	151,949 (21.0)	8462 (8.1)
No. of tetranucleotides (%)	2359 (0.5)	2314 (1.3)	3409 (0.5)	565 (0.5)
No. of pentanucleotides (%)	525 (0.1)	810 (0.5)	905 (0.1)	99 (0.1)
No. of hexanucleotides (%)	698 (0.1)	1274 (0.7)	1573 (0.2)	27 (0.0)
SSR Locator				
No. of dinucleotides (%)	395,436 (82.1)	123,818 (72.4)	565,033 (78.2)	95,062 (91.3)
No. of trinucleotides (%)	82,881 (17.2)	42,773 (25.0)	151,690 (21.0)	8373 (8.0)
No. of tetranucleotides (%)	2335 (0.5)	2288 (1.3)	3384 (0.5)	561 (0.5)
No. of pentanucleotides (%)	516 (0.1)	800 (0.5)	904 (0.1)	97 (0.1)
No. of hexanucleotides (%)	695 (0.1)	1255 (0.7)	1569 (0.2)	27 (0.0)

[a] 6.6 million paired-end reads (2 × 250; 13.2 million total reads); 3.3-Gbp sequence; FASTA file: 3.9 GB.
[b] 10.9 million single-end reads (1 × 100); 1.5-Gbp sequence; FASTA file: 2.2 GB.
[c] 48.5 million paired-end reads (2 × 100; 97 million total reads); 8.7-Gbp sequence; FASTA file: 5.7 GB.
[d] 163,500 reads; 476-Mbp sequence; FASTA file: 445 MB.

comparison to known transposable elements (Meglécz et al., 2014). It can be used both on command-line and through Galaxy (Afgan et al., 2016) and works with RepeatMasker (Tarailo-Graovac and Chen, 2009) and a variety of other NGS tools. Its running time is relatively long (for a 5.7-GB data set, 9.5 h on a high-performance computer), and users cannot change default settings for SSR searches (e.g., specifying different numbers of repeats for different length motifs).

SSR Locator integrates functions of SSR search, frequency of occurrence of motifs, primer design, and PCR simulation against other databases, as well as global alignments and identity and homology searches (da Maia et al., 2008). It executes all the module-calls using a GUI with a built-in menu system suite. However, it requires some file reformatting, which increases computing time. For the HiSeq2 data set, it took 10 min to reformat, and 69 min for the SSR search on a Windows platform (eight Core 3.4-GHz Intel Core i7-2600 CPU, 16 GB RAM).

SSR_pipeline is a command-line program for identifying microsatellites from high-throughput sequencing data using a Python environment (Miller et al., 2013). It detects SSRs in Illumina paired-end reads, with modules for quality filtering and alignment of Illumina raw data. SSR_pipeline can also analyze data from other sequencing platforms, such as 454 and Ion Torrent, by using the SSR detection module independently. However, after 24 h of effort by a biologist proficient in bioinformatics, we could not run test data through SSR_pipeline successfully.

STAMP is an updated package of STADEN (Kraemer et al., 2009) for microsatellite detection and primer design, with comprehensive integration of **Phobos** (Mayer, 2007) for tandem repeat detection and analysis. STAMP uses TROLL (Castelo et al., 2002) for tracing back primer pairs to sequence trace files, Primer3 for interactive design and visualization of primers, and SQLite as a database for storing analysis results. Overall, STAMP is a highly flexible, high-throughput, interactive tool for conventional and multiplex microsatellite marker design, avoiding the generation of redundant markers. However, it is complicated—it requires multiple tool command language modules and preinstallation of the STADEN package, and it is not suited for low-coverage NGS data (Meglécz et al., 2014).

DISCUSSION

***Recommendations for researchers*—**Based on our budget estimates, RAD/GBS and microsatellites are approximately equivalent in cost for genotyping 96 individuals, assuming that NGS data are already available for microsatellite development (Appendix 2). However, RAD/GBS will generate many more loci, but it would be much more economical to add additional individuals if using SSRs. If microsatellites can be developed for free using existing public NGS data, it is worth investigating this option—it can considerably reduce the cost (see companion paper [Hodel et al., 2016], which presents over five million SSR loci that can be used in thousands of plant species). As shown in Table 3, publicly available data sets not designed for microsatellite development can be mined to yield many SSR loci to test. Our review of sequencing platforms and the test data sets we used in our software comparison revealed that read length is not as important as expected. Table 4 indicates that while the longer read lengths associated with MiSeq (2×300 bp) certainly yield more loci than the shorter read lengths of HiSeq (e.g., 1×100 bp), there are plenty of loci detected (typically >100,000) with shorter read lengths. 454 sequencing was once considered essential for microsatellite development, because Illumina reads were too short. Now, many types of Illumina sequencing can generate adequate sequence data for generating loci (Appendix S1). Unless a researcher is multiplexing many different species in a single run, we recommend using Illumina MiSeq for its cost efficiency. As shown in Table 4, the Illumina MiSeq generates ample loci relative to other platforms, and it is cheaper and more time-efficient compared to HiSeq, which requires users to fill all eight lanes before a sequencing run can commence. For the software portion of microsatellite development, we recommend using MISA or Phobos (either alone or as implemented in Geneious).

***The future of SSRs—are they up to the task?*—**Microsatellites still have great applicability due to their high polymorphism, relatively easy scoring, testable neutrality, and Mendelian inheritance (Zane et al., 2002). The use of microsatellites will undoubtedly give way to newer technologies such as RAD/GBS as these approaches find wider application. However, microsatellite markers are valuable tools for several reasons. Many study designs simply do not require the high marker density provided by RAD/GBS and benefit more from the inclusion of large numbers of samples. Furthermore, there are thousands of studies that have employed microsatellite markers, and in many cases, the markers available provided too little information to fully address the authors' hypotheses. For such microsatellite legacy projects, using the same markers as existing data sets is preferred to avoid confounding factors. While microsatellites provide limited information per sample, if the inclusion of many individuals is a priority, microsatellites compare favorably with newer techniques. If transcriptomic data are used to identify microsatellites, it may be possible to perform more rigorous tests of selective neutrality in adjacent coding regions of potential loci. This could allow researchers to know whether they were selecting a locus that is part of (or linked to) a gene under directional selection rather than merely documenting any departures from Hardy–Weinberg equilibrium. Also, the high allelic variation of microsatellites compared to sequence-based markers is optimal for the identification of markers present in small subpopulations of interest (e.g., disease-resistant individuals; Miah et al., 2013). Finally, for projects with limited budgets (e.g., conservation genetic surveys), microsatellites will likely continue to be the most economical option for some time (Jennings et al., 2011). For all of these reasons, microsatellites remain a good choice for many systems and questions—with the proper justification and strong questions/hypotheses, they are still appropriate for use in proposals to the National Science Foundation and other funding sources.

LITERATURE CITED

Afgan, E., D. Baker, M. van den Beek, D. Blankenberg, D. Bouvier, M. Čech, J. Chilton, et al. 2016. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. *Nucleic Acids Research* doi: 10.1093/nar/gkw343.

Andersen, J. C., and N. J. Mills. 2014. iMAST: A novel approach to the development of microsatellite loci using barcoded Illumina libraries. *BMC Genomics* 15: 858.

Andolfatto, P., D. Davison, D. Erezyilmaz, T. T. Hu, J. Mast, T. Sunayama-Morita, and D. L. Stern. 2011. Multiplexed shotgun genotyping for rapid and efficient genetic mapping. *Genome Research* 21: 610–617.

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

Arnold, B., R. B. Corbett-Detig, D. Hartl, and K. Bomblies. 2013. RADseq underestimates diversity and introduces genealogical biases due to nonrandom haplotype sampling. *Molecular Ecology* 22: 3179–3190.

Beck, J. B., and J. C. Semple. 2015. Next-generation sampling: Pairing genomics with herbarium specimens provides species-level signal in *Solidago* (Asteraceae). *Applications in Plant Sciences* 3: 1500014.

Beissinger, T. M., C. N. Hirsch, R. S. Sekhon, J. M. Foerster, J. M. Johnson, G. Muttoni, B. Vaillancourt, et al. 2013. Marker density and read depth for genotyping populations using genotyping-by-sequencing. *Genetics* 193: 1073–1081.

Benson, G. 1999. Tandem repeats finder: A program to analyze DNA sequences. *Nucleic Acids Research* 27: 573–580.

Castelo, A. T., W. Martins, and G. R. Gao. 2002. TROLL—Tandem Repeat Occurrence Locator. *Bioinformatics* 18: 634–636.

Castoe, T. A., A. W. Poole, W. Gu, A. P. De Koning, and J. M. Daza. 2010. Rapid identification of thousands of copperhead snake (*Agkistrodon contortrix*) microsatellite loci from modest amounts of 454 shotgun genome sequence. *Molecular Ecology Resources* 10: 341–347.

Castoe, T. A., A. W. Poole, A. P. J. de Koning, K. L. Jones, D. F. Tomback, S. J. Oyler-McCance, J. A. Fike, et al. 2012. Rapid microsatellite identification from Illumina paired-end genomic sequencing in two birds and a snake. *PLoS ONE* 7: e30953.

Churbanov, A., R. Ryan, N. Hasan, D. Bailey, H. Chen, B. Milligan, and P. Houde. 2012. High SSR: High-throughput SSR characterization and locus development from next-gen sequencing data. *Bioinformatics* 28: 2797–2803.

Cronn, R., B. J. Knaus, A. Liston, P. J. Maughan, M. Parks, J. V. Syring, and J. Udall. 2012. Targeted enrichment strategies for next-generation plant biology. *American Journal of Botany* 99: 291–311.

da Maia, L. C., D. A. Palmieri, V. Q. de Souza, M. M. Kopp, F. I. F. de Carvalho, and A. C. de Oliveria. 2008. SSR Locator: Tool for simple sequence repeat discovery integrated with primer design and PCR simulation. *International Journal of Plant Genomics* 2008: 412696.

DaCosta, J. M., and M. D. Sorenson. 2014. Amplification biases and consistent recovery of loci in a double-digest RAD-seq protocol. *PLoS ONE* 9: e106713.

Davey, J. W., and M. L. Blaxter. 2010. RADSeq: Next-generation population genetics. *Briefings in Functional Genomics* 9: 416–423.

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

Davey, J. W., T. Cezard, P. Fuentes-Utrilla, C. Eland, K. Gharbi, and M. L. Blaxter. 2013. Special features of RAD Sequencing data: Implications for genotyping. *Molecular Ecology* 22: 3151–3164.

Echt, C. S., S. Saha, K. V. Krutovsky, K. Wimalanthan, J. E. Erpelding, C. Liang, and C. D. Nelson. 2011. An annotated genetic map of loblolly pine based on microsatellite and cDNA markers. *BMC Genetics* 12: 17.

Edgar, R. C. 2004. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* 32: 1792–1797.

Ellegren, H. 2000. Microsatellite mutations in the germline: Implications for evolutionary inference. *Trends in Genetics* 16: 551–558.

Ellegren, H. 2004. Microsatellites: Simple sequences with complex evolution. *Nature Reviews. Genetics* 5: 435–445.

Elliott, C. P., N. J. Enright, R. J. N. Allcock, M. G. Gardner, E. Meglecz, J. Anthony, and S. L. Krauss. 2014. Microsatellite markers from the Ion Torrent: A multi-species contrast to 454 shotgun sequencing. *Molecular Ecology Resources* 14: 554–568.

Elshire, R. J., J. C. Glaubitz, Q. Sun, J. A. Poland, K. Kawamoto, E. S. Buckler, and S. E. Mitchell. 2011. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS ONE* 6: e19379.

Esselink, G. D., H. Nybom, and B. Vosman. 2004. Assignment of allelic configuration in polyploids using the MAC-PR (Microsatellite DNA Allele Counting-Peak Ratios) method. *Theoretical and Applied Genetics* 109: 402–408.

Etter, P. D., S. Bassham, P. A. Hohenlohe, E. A. Johnson, and W. A. Cresko. 2011. SNP discovery and genotyping for evolutionary genetics using RAD sequencing. *In* V. Orgogozo and M. V. Rockman [eds.], Methods in molecular biology, vol. 772: Molecular methods for evolutionary genetics, 157–178. Humana Press, New York, New York, USA.

Faircloth, B. C. 2008. MSATCOMMANDER: Detection of microsatellite repeat arrays and automated, locus-specific primer design. *Molecular Ecology Resources* 8: 92–94.

Gardner, M. G., A. J. Fitch, T. Bertozzi, and A. J. Lowe. 2011. Rise of the machines—Recommendations for ecologists when using next generation sequencing for microsatellite development. *Molecular Ecology Resources* 11: 1093–1101.

Gautier, M., K. Gharbi, T. Cezard, J. Foucaud, C. Kerdelhué, P. Pudlo, J.-M. Cornuet, and A. Estoup. 2013. The effect of RAD allele dropout on the estimation of genetic variation within and between populations. *Molecular Ecology* 22: 3165–3178.

Gilmore, B., N. Bassil, A. Nyberg, B. Knaus, D. Smith, D. L. Barney, and K. Hummer. 2013. Microsatellite marker development in peony using next generation sequencing. *Journal of the American Society for Horticultural Science* 138: 64–74.

Glenn, T. 2011. Field guide to next-generation DNA sequencers. *Molecular Ecology* 11: 759–769.

Glenn, T. 2014. 2014 NGS Field Guide: Overview. Available at website http://molecularecologist.com/next-gen-fieldguide-2014/ [accessed 15 February 2016].

Grohme, M. A., R. F. Soler, M. Wink, and M. Frohme. 2013. Microsatellite marker discovery using single molecule real-time circular consensus sequencing on the Pacific Biosciences RS. *BioTechniques* 55: 253–256.

Guichoux, E., L. Lagache, S. Wagner, P. Chaumeil, P. Léger, O. Lepais, C. Lepoittevin, et al. 2011. Current trends in microsatellite genotyping. *Molecular Ecology Resources* 11: 591–611.

Hodel, R. G., M. A. Gitzendanner, C. C. Germain-Aubrey, X. Liu, A. A. Crowl, M. Sun, J. B. Landis, et al. 2016. A new resource for the development of SSR markers: Millions of loci from a thousand plant transcriptomes. *Applications in Plant Sciences* 4: 1600024.

Hoffman, J. I., and W. Amos. 2005. Microsatellite genotyping errors: Detection approaches, common sources and consequences for paternal exclusion. *Molecular Ecology* 14: 599–612.

Hohenlohe, P. A., S. Bassham, P. D. Etter, N. Stiffler, E. A. Johnson, and W. A. Cresko. 2010. Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. *PLoS Genetics* 6: e1000862.

Hou, Y., M. D. Nowak, V. Mirré, C. S. Bjorå, C. Brochmann, and M. Popp. 2015. Thousands of RAD-seq loci fully resolve the phylogeny of the highly disjunct Arctic-Alpine genus *Diapensia* (Diapensiaceae). *PLoS ONE* 10: e0140175.

Huang, H., and L. L. Knowles. 2016. Unforeseen consequences of excluding missing data from next-generation sequences: Simulation study of RAD sequences. *Systematic Biology* 65: 357–365.

Huey, J. A., K. M. Real, P. B. Mather, V. Chand, D. T. Roberts, T. Espinoza, A. McDougall, et al. 2013. Isolation and characterization of 21 polymorphic microsatellite loci in the iconic Australian lungfish, *Neoceratodus forsteri*, using the Ion Torrent next-generation sequencing platform. *Conservation Genetics Resources* 5: 737–740.

Jacobs, D. E., H. J. Dagleish, and C. D. Nelson. 2013. A conceptual framework for restoration of threatened plants: The effective model of American chestnut (*Castanea dentata*) reintroduction. *New Phytologist* 197: 378–393.

Jakobsson, M., M. D. Edge, and N. A. Rosenberg. 2013. The relationship between F_{ST} and the frequency of the most frequent allele. *Genetics* 193: 515–528.

Jennings, T. N., B. J. Knaus, T. D. Mullins, S. M. Haig, and R. C. Cronn. 2011. Multiplexed microsatellite recovery using massively parallel sequencing. *Molecular Ecology Resources* 11: 1060–1067.

Kalia, R. K., M. K. Rai, S. Kalia, R. Singh, and A. K. Dhawan. 2011. Microsatellite markers: An overview of the recent progress in plants. *Euphytica* 177: 309–334.

Kameyama, Y., and A. S. Hirao. 2014. Development and evaluation of microsatellite markers for the gynodioecious shrub *Daphne jezoensis* (Thymelaeaceae). *Applications in Plant Sciences* 2: 1400001.

Kearse, M., R. Moir, A. Wilson, S. Stones-Havas, M. Cheung, S. Sturrock, S. Buxton, et al. 2012. Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* 28: 1647–1649.

Kraemer, L., B. Beszteri, S. Gäbler-Schwarz, C. Held, F. Leese, C. Mayer, K. Pöhlmann, and S. Frickenhaus. 2009. STAMP: Extensions to the STADEN sequence analysis package for high throughput interactive microsatellite maker design. *BMC Bioinformatics* 10: 41.

Landis, J. B., R. D. O'Toole, K. L. Ventura, M. A. Gitzendanner, D. G. Oppenheimer, D. E. Soltis, and P. S. Soltis. 2015. The phenotypic and genetic underpinnings of flower size in Polemoniaceae. *Frontiers in Plant Science* 6: 1144.

Lepais, O., and C. F. E. Bacles. 2011. *De novo* discovery and multiplexed amplification of microsatellite markers for black alder (*Alnus glutinosa*) and related species using SSR-enriched shotgun pyrosequencing. *Journal of Heredity* 102: 627–632.

Li, Y.-C., A. B. Korol, T. Fahima, A. Beiles, and E. Nevo. 2002. Microsatellites: Genomic distribution, putative functions and mutational mechanisms: A review. *Molecular Ecology* 11: 2453–2465.

Liu, M. M., J. W. Davey, R. Banerjee, J. Han, F. Yang, A. Aboobaker, M. L. Blaxter, and A. Davison. 2013. Fine mapping of the pond snail left-right asymmetry (Chirality) locus using RAD-Seq and Fibre-FISH. *PLoS ONE* 8: e71067.

Liu, H., M. Bayer, A. Druka, J. R. Russell, C. A. Hackett, J. Poland, L. Ramsay, P. E. Hedley, and R. Waugh. 2014. An evaluation of genotyping by sequencing (GBS) to map the *Breviaristatum-e* (*ari-e*) locus in cultivated barley. *BMC Genomics* 15: 104.

Lu, F., A. E. Lipka, J. Glaubitz, R. Elshire, J. H. Cherney, M. D. Casler, E. S. Buckler, and D. E. Costich. 2013. Switchgrass genomic diversity, ploidy, and evolution: Novel insights from a network-based SNP discovery protocol. *PLoS Genetics* 9: e1003215.

Mamani, E. M., N. W. Bueno, D. A. Faria, L. M. Guimaaras, D. Lau, A. C. Alfenas, and D. Grattapaglia. 2010. Positioning of the major locus for *Puccinia psidii* rust resistances Ppr1 on the *Eucalyptus* reference map and its validation across unrelated pedigrees. *Tree Genetics & Genomes* 6: 953–962.

Matasci, N., L.-H. Hung, Z. Yan, E. J. Carpenter, N. J. Wickett, S. Mirarab, N. Nguyen, et al. 2014. Data access for the 1,000 plants (1KP) project. *GigaScience* 3: 17.

Mayer, C. 2007. PHOBOS—A tandem repeat search tool for complete genomes. Website http://www.rub.de/spezzoo/cm/cm_phobos.htm [accessed 12 January 2016].

Meglécz, E., N. Pech, A. Gilles, V. Dubut, P. Hingamp, A. Trilles, R. Grenier, and J. F. Martin. 2014. QDD version 3.1: A user-friendly computer program for microsatellite selection and primer design revisited: Experimental validation of variables determining genotyping success rate. *Molecular Ecology Resources* 14: 1302–1313.

Miah, G., M. Y. Rafii, M. R. Ismail, A. B. Puteh, H. A. Rahim, K. N. Islam, and M. A. Latif. 2013. A review of microsatellite markers and their applications in rice breeding programs to improve blast disease resistance. *International Journal of Molecular Sciences* 14: 22499–22528.

Miller, M. P., B. J. Knaus, T. D. Mullins, and S. M. Haig. 2013. SSR_ pipeline: A bioinformatic infrastructure for identifying microsatellites from paired-end Illumina high-throughput DNA sequencing data. *Journal of Heredity* 104: 881–885.

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

Nunome, T., S. Negoro, K. Miyatake, H. Yamaguchi, and H. Fukuoka. 2006. A protocol for the construction of microsatellite enriched genomic library. *Plant Molecular Biology Reporter* 24: 305–312.

Poland, J. A., and T. W. Rife. 2012. Genotyping-by-sequencing for plant breeding and genetics. *Plant Genome* 5: 92–102.

Poland, J., J. Endelman, J. Dawson, J. Rutkoski, S. Wu, Y. Manes, S. Dreisigacker, et al. 2012. Genomic selection in wheat breeding using genotyping-by-sequencing. *Plant Genome* 5: 103–113.

Rafalski, A. 2002. Applications of single nucleotide polymorphisms in crop genetics. *Current Opinion in Plant Biology* 5: 94–100.

Ritschel, P. S., T. C. de Lima Lins, R. L. Tristan, G. S. C. Buso, J. A. Buso, and M. E. Ferreira. 2004. Development of microsatellite markers from an enriched genomic library for genetic analysis of melon (*Cucumis melo* L.). *BMC Plant Biology* 4: 9.

Rowe, H. C., S. Renaut, and A. Guggisberg. 2011. RAD in the realm of next-generation sequencing technologies. *Molecular Ecology* 20: 3499–3502.

Rozen, S., and H. Skaletsky. 1999. Primer3 on the WWW for general users and for biologist programmers. *In* S. Misener and S. A. Krawetz [eds.], Methods in molecular biology, vol. 132: Bioinformatics methods and protocols, 365–386. Humana Press, Totowa, New Jersey, USA.

Schoebel, C. N., S. Brodbeck, D. Buehler, C. Cornejo, J. Gajurel, H. Hartikainen, D. Keller, et al. 2013. Lessons learned from microsatellite development for nonmodel organisms using 454 pyrosequencing. *Journal of Evolutionary Biology* 26: 600–611.

Selkoe, K. A., and R. J. Toonen. 2006. Microsatellites for ecologists: A practical guide to using and evaluating microsatellite markers. *Ecology Letters* 9: 615–629.

Serikawa, T., T. Kuramoto, P. Hilbert, M. Mori, J. Yamada, C. J. Dubay, K. Lindpainter, et al. 1992. Rat gene mapping using PCR-analyzed microsatellites. *Genetics* 131: 701–721.

Slatkin, M. 1995. A measure of population subdivision based on microsatellite allele frequencies. *Genetics* 139: 457–462.

Smith, A. M., L. E. Heisler, R. P. St. Onge, E. Farias-Hesson, I. M. Wallace, J. Bodeau, A. N. Harris, et al. 2010. Highly-multiplexed barcode sequencing: An efficient method for parallel analysis of pooled samples. *Nucleic Acids Research* 38: e142.

Soltis, D. E., P. S. Soltis, and J. J. Doyle [eds.]. 1998. Molecular systematics of plants II. Springer, Boston, Massachusetts, USA.

Sonah, H., M. Bastien, E. Iquira, A. Tardivel, G. Légaré, B. Boyle, É. Normandeau, et al. 2013. An improved genotyping by sequencing (GBS) approach offering increased versatility and efficiency of SNP discovery and genotyping. *PLoS ONE* 8: e54603.

Spindel, J., M. Wright, C. Chen, J. Cobb, J. Gage, S. Harrington, M. Lorieux, et al. 2013. Bridging the genotyping gap: Using genotyping by sequencing (GBS) to add high-density SNP markers and new value to traditional bi-parental mapping and breeding populations. *Theoretical and Applied Genetics* 126: 2699–2716.

Tarailo-Graovac, M., and N. Chen. 2009. Using RepeatMasker to identify repetitive elements in genomic sequences. *Current Protocols in Bioinformatics* 4: 10.

Thiel, T., W. Michalek, R. K. Varshney, and A. Graner. 2003. Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (*Hordeum vulgare* L.). *Theoretical and Applied Genetics* 106: 411–422.

Viard, F., P. Franck, M. P. Dubois, A. Estoup, and P. Jarne. 1998. Variation of microsatellite size homoplasy across electromorphs, loci, and populations in three invertebrate species. *Journal of Molecular Evolution* 47: 42–51.

Wandeler, P., P. E. A. Hoeck, and L. F. Keller. 2007. Back to the future: Museum specimens in population genetics. *Trends in Ecology & Evolution* 22: 634–642.

Wang, X., P. Lu, and Z. Luo. 2013. GMATo: A novel tool for the identification and analysis of microsatellites in large genomes. *Bioinformation* 9: 541–544.

Wainwright, B. J., I. S. Arlyza, and S. A. Karl. 2013. Isolation and characterization of twenty-one polymorphic microsatellite loci for *Polycarpa aurata* using third generation sequencing. *Conservation Genetics Resources* 5: 671–673.

Wei, N. A., J. B. Bemmels, and C. W. Dick. 2014. The effects of read length, quality and quantity on microsatellite discovery and primer development: from Illumina to PacBio. *Molecular Ecology Resources* 14: 953–965.

Whitlock, M. C. 2011. G'_{ST} and D do not replace F_{ST}. *Molecular Ecology* 20: 1083–1091.

Xu, P., S. Xu, X. Wu, Y. Tao, B. Wang, S. Wang, D. Qin, et al. 2014. Population genomic analyses from low-coverage RAD-Seq data: A case study on the non-model cucurbit bottle gourd. *Plant Journal* 77: 430–442.

Zalapa, J. E., H. Cuevas, H. Zhu, S. Steffan, D. Senalik, E. Zelden, B. McCown, R. Harbut, and P. Simon. 2012. Using next-generation sequencing approaches to isolate simple sequence repeat (SSR) loci in the plant sciences. *American Journal of Botany* 99: 193–208.

Zane, L., L. Bargelloni, and T. Patarnello. 2002. Strategies for microsatellite isolation: A review. *Molecular Ecology* 11: 1–16.

Appendix 1. Glossary of terms.

Flanking regions: Sequences on either side of the microsatellite repeat. These regions are where the primers anneal for microsatellite amplification. Uniqueness of flanking sequences for each locus is assumed. High GC content is recommended to improve the stability of primer sites.

Genotyping by sequencing (GBS) and restriction site associated DNA sequencing (RAD-Seq): This suite of methods uses restriction enzymes to reduce genome complexity and NGS to sequence thousands of single-nucleotide polymorphisms (SNPs) for hundreds to thousands of individuals in a multiplexed reaction. These methods have several variants, but collectively they represent an alternative to microsatellites and may be replacing them given the massive amounts of data they generate for comparable costs.

Hybridization-based enrichment: Microsatellite repeat–containing probes are attached to magnetic beads or nylon membranes, and hybridization between probes and target DNA is used to capture DNA fragments containing microsatellites.

Microsatellite development: The process of developing primers to amplify microsatellite loci. Source data can be obtained from NGS data, repeat-enriched clones generated from genomic libraries, or by screening sequences in databases.

Multiplex sequencing: This technique simultaneously sequences numerous samples in a single sequencing run. Samples are given diagnostic sequence tags and can then be mixed, sequenced together, and bioinformatically separated for data analysis.

Neutral vs. non-neutral: Neutral loci are not under the influence of natural selection, and patterns of variation reflect interactions among mutation, drift, mating system, and migration. Non-neutral loci are subject to selection—either directly or indirectly when the locus in question is linked to a region of the genome under selection.

Next-generation sequencing (NGS): Newer DNA sequencing technologies (e.g., Illumina, Roche 454, Pacific Biosciences) that generate vastly more sequence than Sanger sequencing methods, greatly increasing the amount of data obtained, while reducing the per-base cost of DNA sequencing.

Transcriptome: All the messenger RNA (mRNA) molecules expressed by a cell, tissue, or organism. The study of transcriptomes allows analyses of gene expression as well as variants of mRNA arising through alternative splicing, RNA editing, or alternative transcription initiation and termination sites.

Appendix 2. Sample budgets for genotyping 96 individuals using microsatellites or RAD/GBS. All costs are expressed as 2016 US dollars.

Costs associated with developing simple sequence repeat (SSR) loci include next-generation sequencing (NGS), fluorescently labeled primers for loci that pass initial criteria, PCR reagents, gel electrophoresis, and the apparatus for detecting fluorescent peaks. Although there are several methods for developing microsatellite loci, we present one commonly used method to simplify our comparison of microsatellite and RAD/GBS budgets. Based on personal experience, we estimate the total cost for undertaking a microsatellite genotyping project of 96 individuals for 12–15 loci to be approximately $4100 (Table A1). This estimate assumes that NGS data have already been generated—if this is not the case, the initial cost could increase considerably (by up to $1000). The costs we include fall into three categories: initial screening with unlabeled primers, screening labeled primers, and genotyping. For initial screening, a reasonable starting point is to order unlabeled primers for 48 loci and to screen these loci for amplification using eight individuals. Each primer pair costs $12, for a total of $576 (48 primer pairs at $12 each). Additionally, approximately two QIAGEN PCR multiplex kits are needed for the initial screening step, which cost $540 (the kits are $270 each). We estimate the total cost for the initial screening to be $828. We assume that half the loci are rejected in the initial screening, due to no amplification, multiple bands, amplification in only a few individuals, or some other amplification issues. Thus, there are 24 potential loci that move to the screening with labeled primers step. In this step, one primer per locus is replaced with a labeled primer that costs $80, for a total of $1920 (24 potential loci to be tested at $80 each). Additionally, another PCR kit will be needed for this step, adding another $270. As these 24 loci are screened against eight samples, it will require the genotyping of two plates ($200). Note that it is advantageous to order labeled primers in two batches, to make it easier to optimize the assignment of different dyes to loci of similar size, making it possible to include at least four loci multiplexed in one well for genotyping. Also, these samples can be considered replicates to assess error rates in subsequent analyses. The total cost for the labeled primer screening step is $2390 ($1920 + $270 + $200). Once again, we assume that half the loci are lost in the second step, leaving 12 good loci to use to genotype all the individuals. Assuming that it is still possible to multiplex four loci in a well, three plates could be used to genotype all 96 individuals (adding $300 to the research budget, assuming $100 per plate for genotyping costs). Another PCR kit would also be required for the genotyping step as well, adding another $270 to the budget. We consider that 50% loss of loci at each step is rather conservative; therefore, we think it is fair to assume that 12–15 loci could be developed for these costs (instead of only 12 loci). It is important to note that careful planning and judicious multiplexing may greatly reduce costs. Another important consideration for microsatellites is that it is generally quite economical to add additional samples. For instance, once the markers are developed, the only costs are reagents, consumables, and lane charges, which are less than $2 per sample per PCR. Thus, doubling the number of samples to 192 could be accomplished for approximately an additional $800.

RAD/GBS costs depend on the type of digestion and number of samples. Typical costs associated with RAD/GBS include purchasing restriction enzymes and other reagents, sample quantification, sample quality control, size selection (not used in all methods), and sequencing. Based on personal experience, the calculator provided by the Oregon State Center for Genome Research and Biocomputing (CGRB; http://hts2.cgrb.oregonstate.edu/calc/gbs/ [accessed 21 February 2016]), and a pricing quote from the University of Texas Genomic Sequencing and Analysis Facility (GSAF) website (https://wikis.utexas.edu/display/GSAF/Pricing [accessed 14 May 2016]), we estimate the cost of RAD/GBS with 96 samples to range from approximately $3400 to $5000. For the lower figure, we based our estimate on the public price quote from the University of Texas GSAF. This website quotes that double digest RAD prep costs $31.92 per sample for 96 samples, plus a fixed cost of $340, which yields a grand total of $3404.32. We arrived at the higher figure by assuming it is necessary to pay for a double digest ($1261), digest optimization ($539), 10 QC Bioanalyzer Chips ($1010), Qubit quantification ($400), one lane of Illumina HiSeq 3000 (1 × 150; $1225), dsDNA Fluorophore quantification ($69), and reagent cost of approximately $500; these prices are from the Oregon State University CGRB (Table A1).

Based on these budgets, the cost of using RAD/GBS to genotype thousands of loci for 96 individuals is comparable to developing 12–15 microsatellite loci to genotype the 96 individuals. RAD/GBS project costs ranged from $3400–$5000, and microsatellite development and genotyping costs are at least $4100; this number will increase if an NGS run is necessary to generate sequence data to mine for loci. We did not include costs such as DNA extraction and gel electrophoresis,

because both microsatellites and RAD/GBS would have similar expenses in those categories. One additional consideration is that microsatellite development is labor-intensive—for some projects, it will make more sense to pay for RAD/GBS to save the time that it would take to develop loci. Unlike with microsatellites, doubling the number of individuals in a RAD/GBS study would nearly double the project budget. The one key cost-saving feature of RAD/GBS is that multiplexing individuals makes the NGS costs affordable; however, this is the reason why it is very costly to add one additional sample to a study.

TABLE A1. Costs associated with microsatellite development for 12–15 loci and RAD/GBS costs to generate thousands of loci. Both budgets assume 96 individuals are included in the study.

Item	Base cost	Quantity	Total cost
Microsatellites			
QIAGEN PCR multiplex kit	$270	4	$1080
Unlabeled primer pairs	$6	48	$288
Labeled primer (single)	$80	24	$1920
Genotyping one plate	$100	4	$400
Total			**$3688**
RAD/GBS (high estimate)			
Double digest	$1261	1	$1261
Digest optimization	$539	1	$539
QC Bioanalyzer chips	$101	10	$1010
Qubit quantification	$400	1	$400
Illumina HiSeq lane	$1225	1	$1225
dsDNA Fluorophore quantification	$69	1	$69
Reagents	$500	1	$500
Total			**$5004**
RAD/GBS (low estimate)			
Per sample cost	$31.92	96	$3064.32
Fixed cost	$340	1	$340
Total			**$3404.32**