AN EMPIRICAL REVIEW: CHARACTERISTICS OF PLANT MICROSATellite MARKERS THAT CONFER HIGHER LEVELS OF GENETIC VARIATION

Benjamin J. Merritt, Theresa M. Culley, Alina Avanesyan, Richard Stokes, and Jessica Brzyski

During microsatellite marker development, researchers must choose from a pool of possible primer pairs to further test in their species of interest. In many cases, the goal is maximizing detectable levels of genetic variation. To guide researchers and determine which markers are associated with higher levels of genetic variation, we conducted a literature review based on 6782 genomic microsatellite markers published from 1997–2012. We examined relationships between heterozygosity ($H_e$ or $H_o$) or allele number ($A$) with the following marker characteristics: repeat type, motif length, motif region, repeat frequency, and microsatellite size. Variation across taxonomic groups was also analyzed. There were significant differences between imperfect and perfect repeat types in $A$ and $H_e$. Dinucleotide motifs exhibited significantly higher $A$, $H_e$, and $H_o$ than most other motifs. Repeat frequency and motif region were positively correlated with $A$, $H_e$, and $H_o$, but correlations with microsatellite size were minimal. Higher taxonomic groups were disproportionately represented in the literature and showed little consistency. In conclusion, researchers should carefully consider marker characteristics so they can be tailored to the desired application. If researchers aim to target high genetic variation, dinucleotide motif lengths with large repeat frequencies may be best.

Key words: genetic variation; heterozygosity; microsatellite; motif; repeat frequency; taxonomy.

For many researchers, microsatellites continue to be the marker of choice for surveys of genetic diversity and structure, as well as paternity analysis and mating system estimates in which codominance is essential. Microsatellites, also known as simple sequence repeats (SSRs) or short tandem repeats (STRs), are typically defined as repeated sequences of one to six bases found throughout the nuclear and plastid genomes of eukaryotes (e.g., Zane et al., 2002; Buschiazzo and Gemmell, 2006; Wheeler et al., 2014). Despite the many benefits of these markers (see Estoup and Angers, 1998; Goldstein and Schlötterer, 1999; Selkoe and Toonen, 2006), a disadvantage in developing genomic markers is that in many cases, microsatellite primers must be developed de novo for a species, especially if primers are not available for testing from closely related species or genera. Even then, primers from related taxa may not be conserved (Rubinsztejn et al., 1995; Primmer et al., 1996; Whitton et al., 1997; Morin et al., 1998), often requiring de novo development on a species-by-species basis. Traditional methods of microsatellite marker development involve construction of a genomic library through enrichment for microsatellite repeats, cloning, plasmid isolation, and Sanger sequencing (Zane et al., 2002). Although this process can generate several hundred sequences, only a small subset are usually acceptable for subsequent primer design and evaluation because they must contain a desired repeat, be of the appropriate size (usually 100–300 bp), and have suitable room for primer design in the regions flanking the repeated motif (Squirrell et al., 2003). More recently, next-generation sequencing (NGS) technology has dramatically increased the yield of potential microsatellite primer pairs, generating thousands of individual reads (Ekkblom and Galindo, 2011; Hoffman and Nichols, 2011), of which at least 2000 primer pairs may be suitable for further testing (Abdelkrim et al., 2009). Consequently, researchers using either traditional or NGS approaches are eventually faced with an array of primer pairs from which a subset must be selected for further testing in the focal species. How does an investigator decide which primers to choose for further development?

Although there are no commonly accepted criteria for selecting these primer pairs, investigators often choose certain markers based on specific characteristics, which are first described as follows. The nucleotide composition of the repeated sequence is called the motif (Abdelkrim et al., 2009), which can be further described by the motif length, also known as the repeat length (Weber, 1990; Scribner and Pearce, 2000) or sometimes repeat unit (Urquhart et al., 1994) (italicized terms are defined in Appendix 1). The motif length reflects the number of bases in the motif that are repeated [e.g., mononucleotide: (T)$_n$, dinucleotide: (TA)$_n$, trinucleotide: (CGG)$_n$, tetranucleotide: (GAAT)$_n$, pentanucleotide: (GATT)$_n$, and hexanucleotide: (CCGGTA)$_n$]. The

1 Manuscript received 13 March 2015; revision accepted 8 July 2015.

2 Department of Biological Science, University of Cincinnati, 614 Rieveschl Hall, Cincinnati, Ohio 45221-0006 USA; 3 Iowa State University, 1317 Illinois Avenue, Ames, Iowa 50014 USA; 4 University of Illinois at Springfield, One University Plaza, MS HSB 224, Springfield, Illinois 62703-5407 USA; and 5 Department of Biology, Seton Hill University, 1 Seton Hill Drive, Greensburg, Pennsylvania 15601 USA.

doi:10.3732/apps.1500025

Applications in Plant Sciences 2015 3(8): 1500025; http://www.bioone.org/loi/apps © 2015 Merritt et al. Published by the Botanical Society of America. This work is licensed under a Creative Commons Attribution License (CC-BY-NC-SA).
number of times that such a motif appears \( (n) \) is known as the \textit{repeat frequency} or repeat array (Scribner and Pearce, 2000). Multiplying the repeat frequency by the number of base pairs in the motif gives the \textit{motif region} length or motif size range. These motifs may occur in different types of arrangements, often referred to as \textit{motif type}, also known as the motif contiguity (Scribner and Pearce, 2000), repeat pattern, or purity of length (Buschiazzo and Gemmell, 2006). Motif types consist of the following: (1) \textit{perfect repeats} (Estoup and Angers, 1998; Scribner and Pearce, 2000; Bhargava and Fuentes, 2010), also called simple (Levinson and Gutman, 1987) or pure repeats (Rosenbaum and Deinard, 1988; Buschiazzo and Gemmell, 2006), such as \((CA)_n \) or \((GTAG)_n \); (2) \textit{compound repeats}, which are composed of two or more successive sets of perfect repeats, such as \((AT)_n (GTC)_n \) (Weber, 1990; Estoup and Angers, 1998; Bhargava and Fuentes, 2010); and (3) \textit{interrupted repeats}, sometimes called \textit{imperfect repeats} (Estoup and Angers, 1998; Scribner and Pearce, 2000), which contain an intervening, nonrepeat sequence between two or more perfect or compound repeats, e.g., \((TC)_n CTAG(CCG)_n \).

It has been suggested that investigators faced with an array of possible primer pairs should select those associated with dinucleotide repeats over more elaborate motif lengths (tris-, tetra-, or pentanucleotide motifs) to ensure higher levels of genetic variation (Levinson and Gutman, 1987; Grist et al., 1993; Chakraborty et al., 1997; Sup Lee et al., 1999; Ellegren, 2000, 2004). In fact, the majority of microsatellite markers (48–67%) found in many species are dinucleotide repeats, but these are less frequent in coding regions (Li et al., 2002). Trinucleotide and hexanucleotide repeats are thought to be more common in coding regions because they do not cause a frameshift (Toth et al., 2000; Rungis et al., 2004; Russell et al., 2004; Chabane et al., 2005; Woodhead et al., 2005; Martin et al., 2010). For these reasons, we limited the database to only include genomic SSRs that are assumed to be under neutral selection. Furthermore, the removal and exclusion of genic SSRs included many agricultural crops, which are traditionally inbred beyond what is expected in natural populations.

In cases in which statistics from multiple populations were reported, we selected the single population with the largest sample size to represent genetic diversity of that study, instead of using mean values calculated across populations. This was done to maintain consistency across papers (e.g., compared to studies with only single population screenings) and to best represent the population-level variation present in the species. In instances where multiple populations with the same sample size were reported, one population was selected at random to include in the database. Although recent studies suggesting ideal characteristics of microsatellite markers have only investigated markers with a minimum repeat frequency (e.g., \( > 6 \), \( > 10 \), or \( > 20 \); Weber, 1990; Morgante and Olivieri, 1993; Wang et al., 1994), here we did not discriminate against repeat frequency, so as to incorporate the widest possible breadth and depth of markers developed thus far. We also included both monomorphic and polymorphic markers in the database; monomorphic markers thus served as the basis for comparison of genetic variability values.

The primer information within the database for each locus consisted of the following information whenever possible: the reported locus name or merit ID (those entries without either of these criteria were assigned a unique number), the primer motif, the number of alleles \((A)\) and levels of expected and observed heterozygosity \((H_e \text{ and } H_o)\). We were interested in the following questions:

1. Are different motif types (perfect vs. imperfect) associated with different levels of genetic variation?
2. Are smaller motif lengths (di-, tri-, etc.) associated with greater levels of genetic variation?
3. Is a higher repeat frequency or larger motif region associated with greater levels of genetic variation?
4. Is there a relationship between fragment size and levels of genetic variation?
5. In utilizing such a unique data set, are certain taxonomic groups disproportionally represented in the microsatellite primer development literature? Are there any trends in levels of genetic variation as revealed by microsatellite markers among these taxonomic groups?

**MATERIALS AND METHODS**

**Database compilation**—To analyze genomic microsatellites from all plants, including algae, fungi, and both flowering and nonflowering species, we focused on journals in which such markers are usually reported. The database was constructed from predominantly primer note articles published in \textit{Molecular Ecology}, \textit{Molecular Ecology Notes}, \textit{Molecular Ecology Resources}, and \textit{American Journal of Botany} (AJB). Data were obtained in the case of \textit{Molecular Ecology} and associated publications, microsatellite data were obtained directly for the years 1997 to early 2009 from the \textit{Molecular Ecology Resources} online database (http://tomato.bio.trinity.edu), where authors are required to submit microsatellite primer information as a condition of publication. In these cases, all entries were screened to confirm plant species. The remainder of \textit{Molecular Ecology} papers published in 2000–2012 as well as all AJB papers from 1996–2012 were screened manually using the Scopus search engine (Elsevier; http://www.scopus.com) by searching for the key word “microsatellite” in titles, abstracts, and key words while excluding the words “Animal,” “Animals,” or “Aves.” All citations from the manually compiled papers were then exported to a database in EndNote (Mendeley Ltd.; https://www.mendeley.com); the microsatellite primer information and measures of population-level genetic variation (i.e., population screenings) were copied from their appropriate tables within each paper and incorporated into a Microsoft Excel (2007) database together with data previously acquired from the \textit{Molecular Ecology Resources} online database.

Many studies published more recently have embraced NGS technologies in SSR development, pursuing markers that can be mined from publicly available data sets (e.g., the National Center for Biotechnology Information’s [NCBI] GenBank) of genic regions compiled using expressed sequence tags (ESTs). While this approach can generate thousands of putative markers, they are limited to transcribed regions that are presumably under selection, and may exhibit reduced polymorphism compared with genomic SSRs (Cho et al., 2000; Scott et al., 2000; Eujayl et al., 2001; Runigs et al., 2004; Russell et al., 2004; Chabane et al., 2005; Woodhead et al., 2005; Martin et al., 2010). For these reasons, we limited the database to only include genomic SSRs that are assumed to be under neutral selection. Furthermore, the removal and exclusion of genic SSRs included many agricultural crops, which are traditionally inbred beyond what is expected in natural populations.

**Database modifications prior to statistical analysis**—Primers were classified by motif type, as either perfect or imperfect based upon the nature or contiguity of the reported repeat motifs. Perfect motifs were simple sequence repeats (SSRs) ranging from two to six nucleotides [e.g., (AC)\(_n\), or ATTCGG\(_n\)], Imperfect repeats were classified as compound or interrupted motifs, such as (ATG)\(_n\)(AC)\(_n\) or (CT\(_n\)(CA)\(_n\)(CT)\(_n\). The perfect motifs were further classified into different motif lengths, based on the number of nucleotides within the repetitive sequence, from two to six. The imperfect motifs, while included in their own separate bin in this study and reported for comparison purposes, were not statistically analyzed in comparison with other SSR motifs because they are more complex in composition and contain greater size ranges. Repeat frequency was extracted when available for each reported marker. The motif region was also calculated for each of the perfect motifs that also included a repeat frequency for each locus. In papers in

http://www.bioone.org/loi/apps
Fig. 1. Database compilation/modification workflow. This workflow depicts the general origins, compilations, and modifications of each individual entry into the database with a general “binning” of different variables based on the nature of the measurements and the appropriate statistical analyses used.
which a range was reported for the microsatellite size, the minimum, maximum, and mean values of that range were obtained. The lengths of the reported forward and reverse primers (excluding fluorescent label tags) were also calculated. Given that there was very little variation in primer lengths (which only varied within a few base pairs), the data and associated analyses of primer length are not included here (but are available upon request).

Finally, major taxonomic levels from family up to kingdom were incorporated for each species and locus, with a variety of sources used to place taxa accordingly and to update older family classifications as needed (Encyclopedia of Life [http://eol.org/], ITIS [http://www.itis.gov/]; Stevens, 2001; Mabberley, 2008; The Plant List, 2013; Guiry and Guiry, 2014; Index Fungorum, 2014). A subset of these taxonomic levels were grouped along mono- and dicotyledonous lines and then also in major categories to encapsulate the breadth of divergence in the phylum Tracheophyta. These taxonomic groups include gymnosperms, Nymphaeales, Austrobaileyales, magnoliids, true eudicots, rosids, and asterids. Mesquite 3.0 build 644 was used to build a cladogram based on the taxonomic tree from the Angiosperm Phylogeny Group (Stevens, 2001).

**Duplicate entries**—As with any database compilation, duplicate entries had to be addressed. In most cases, straightforward repeated entries were removed. There were, however, a handful of cases of other types of repeated entries that were dealt with individually. Primers were only kept in the database if they included genetic variation parameters for the original species for which they were designed. Therefore, nonspecific primers designed in similarly related species were removed, with the following exceptions. (1) There were three publications (34 entries) in which primer pairs were designed using DNA from two species and the population screening information was provided for distinct populations of each species separately. These were maintained in the database because the primers were effectively species-specific in their design process. (2) In addition, 24 primer pairs revealed multiple loci of amplification in their appropriate species and the population screening information was maintained for each locus as separate primers. (3) Finally, there were 15 entries in which the primer pairs matched but the motif was different. These were all within-species duplicates and were maintained as separate markers. It should be noted that in all cases listed above, alternative permutations, such as reverse complements, were not considered because of the complexity of the data.

The data set with modifications as described above is available on the Dryad Digital Repository (http://dx.doi.org/10.5061/dryad.7gr39; Merritt et al., 2015).

**Statistical analysis**—SAS/STAT version 9.4 (SAS Institute, Cary, North Carolina, USA) was used to statistically analyze associations across different motif lengths, motif types, microsatellite sizes, repeat frequencies, motif regions, and taxonomic groups with levels of genetic variation, quantified as A, \(H_e\), \(H_o\), and \(e\). A Spearman’s rank correlation was used to compare all noncategorical data to identify associations between levels of genetic diversity and microsatellite marker traits. Contingency tables and Fisher’s exact or chi-square tests were used to identify whether certain motif lengths were associated with morphomorphic markers. Because the data set exhibited significant deviations from normality in both inspection of quantile-quantile plots and according to the Kolmogorov–Smirnov test, the Kruskal–Wallis test (PROC NPAR1WAY) was used for categorical comparisons. Preliminary analysis used ANOVAs (PROC GLM) with type III sums of squares because this parametric test is generally robust and resistant to deviations from normality; although these tests are not reported here (available upon request), they were in agreement with the results of the Kruskal–Wallis tests.

The posthoc Dwass–Steel–Critchlow–Fligner pairwise comparison (DSCF, a nonparametric equivalent of Tukey’s honest significant difference) test was used to examine differences in levels of genetic variation between each of the groups tested that involve motif lengths.

### RESULTS

**Trait correlations**—A was strongly correlated with \(H_e\) (Spearman’s rank correlation coefficient, \(r_s = 0.835, P < 0.0001\); see Table 1) and \(H_o\) (\(r_s = 0.530, P < 0.0001\)). Furthermore, \(H_e\) and \(H_o\) were significantly correlated with one another (\(r_s = 0.651, P < 0.0001\)). There were strong, positive correlations between A and repeat frequency (\(r_s = 0.431, P < 0.0001\)) and A and motif region (\(r_s = 0.413, P < 0.0001\)). No significant correlation was found between A and mean microsatellite size (\(r_s = 0.00889, P = 0.467\)); however, A was inversely correlated with minimum microsatellite size (\(r_s = -0.0769, P < 0.0001\)) and positively correlated with maximum microsatellite size (\(r_s = 0.127, P < 0.0001\)). There were no significant correlations with mean microsatellite size and \(H_e\) (\(r_s = -0.0153, P = 0.2226\)) or \(H_o\) (\(r_s = -0.0233, P = 0.0714\)). A slight but significant inverse correlation was found between minimum microsatellite size and both \(H_e\) (\(r_s = -0.0625, P < 0.0001\)) and \(H_o\) (\(r_s = -0.0581, P < 0.0001\)); however, maximum microsatellite size was positively correlated with both \(H_e\) (\(r_s = 0.101, P < 0.0001\)) and \(H_o\) (\(r_s = 0.0306, P = 0.0337\)). Repeat frequency was significantly correlated with \(H_e\) (\(r_s = 0.395, P < 0.0001\)) and \(H_o\) (\(r_s = 0.246, P < 0.0001\)), but there was no significant correlation with mean microsatellite size (\(r_s = 0.00817, P = 0.5864\)).

**Motif analysis**—In analyzing the specific perfect motifs and the imperfect motifs, there were approximately 3061 different motifs reported in the database out of 6782 entries (this estimate refers to unique motifs with differing repeat frequencies and does not take into consideration alternative permutations described below). In the case of the unique dinucleotide repeats, the most abundantly reported motif was \(GA_n\) (including complementary, reverse, and reverse-complementary permutations: CT, AG, and TC in descending order of frequency), accounting for approximately 34% of all motifs in the data set and 66% of all dinucleotide repeats. The second most abundant dinucleotide motif was \(CA_n\) (including the reverse, reverse complement, and complementary permutations: AC, GT, and TG) accounting for 15% of all motifs and 30% of all dinucleotide repeats. Of the trinucleotide repeats, the top three most commonly reported were \(CTT_n\) (including AAG, GAA, and TTC; 31.6% of trinucleotide repeats and 4.01% of all motifs), \(CAA_n\) (ACC, GTT, and TGG; 11.3% of trinucleotide repeats and 1.43% of all motifs),

### Table 1: Spearman’s rank correlation matrix comparing levels of genetic variation and marker traits. The upper right side of the table includes noncategorical data for each pairwise comparison in parentheses. Significant (\(P < 0.05\)) correlation coefficients are in bold.

|       | \(A\)  | \(H_e\) | \(H_o\) | Repeat frequency | Motif region | Mean size |
|-------|-------|--------|--------|-----------------|--------------|----------|
| \(A\) | 0.834 | 0.530  | 0.431  | 0.413           | 0.00889      |
| \(H_e\)| \(<0.0001 (6336)\) | \(<0.0001 (750)\) | 0.651  | 0.388           | 0.0153      |
| \(H_o\)| \(<0.0001 (6006)\) | \(<0.0001 (4250)\) | \(<0.0001 (4017)\) | 0.246 | 0.361 |
| Repeat frequency | \(<0.0001 (4437)\) | \(<0.0001 (4250)\) | \(<0.0001 (4017)\) | \(<0.0001 (4485)\) | \(0.00817\) |
| Motif region | \(<0.0001 (4437)\) | \(<0.0001 (4250)\) | \(<0.0001 (4017)\) | \(<0.0001 (4485)\) | \(0.0396\) |
| Mean size | 0.467 (6687) | 0.2226 (6327) | 0.0714 (6016) | 0.5864 (4434) | 0.0084 (4434) |

**Note:** \(A\) = number of alleles; \(H_e\) = expected heterozygosity; \(H_o\) = observed heterozygosity.
and GAT<sub>n</sub> (ATC, CTA, and TAG; 6.5% of trinucleotide repeats and 0.826% of all motifs).

Compared with imperfect motifs, perfect motifs as a group exhibited significantly higher levels of A and H<sub>e</sub> (H<sub>e</sub> = 4.36 and 5.06; P = 0.037 and 0.025, respectively; see Table 2); however, there were no significant differences in H<sub>o</sub> (H<sub>o</sub> = 0.04; P = 0.8513). Within perfect motifs, motif lengths differed significantly from one another for A and H<sub>e</sub> (H = 107.89, 132.96, and 82.08; P < 0.0001, respectively; see Table 2). The dinucleotide repeat motifs exhibited significantly higher H<sub>e</sub> than any other motif length, and significantly higher A and H<sub>e</sub> than the tri-, tetra-, and pentanucleotide repeats (see Table 2, Figs. 2 and 3). Although these significant differences could be a function of the different sample sizes within each motif length group, this is unlikely as the tests incorporate sample size in the calculation.

**Microsatellite characteristics**—The mean, minimum, and maximum microsatellite sizes were significantly lower in perfect motif types compared to imperfect motifs (see Table 2). There was a significant difference among the motif lengths in the mean microsatellite size range, with the general trend of size increasing with the number of nucleotides present in the motif (H = 39.6, P < 0.0001; see Table 2). Within perfect motifs, the variation in minimum, mean, and maximum microsatellite sizes was similar across the different motif lengths with respect to magnitude and direction; therefore, only the mean microsatellite size is reported in Table 2.

The motif region significantly differed among motif lengths (H = 28.4, P < 0.0001), but there was no consistent trend or relationship across the motif lengths (see Table 2, Fig. 4). Repeat frequencies across the different motif lengths showed very strong significant differences between groups, exhibiting an

### Table 2. Marker comparisons of motif types and motif lengths. Each major row includes the mean and number of entries (n) within each category with Kruskal–Wallis statistics (H and the corresponding P value) for comparisons of different major microsatellite characteristic groupings (e.g., motif type and motif length). For each major comparison, levels of genetic variation are included along with repeat frequency, motif region, and mean size. Significant values (P < 0.05) are shown in bold.

| Variable                  | Motif type                  | Motif length<sup>a</sup> | 2  | 3  | 4  | 5  | 6  |
|---------------------------|-----------------------------|---------------------------|----|----|----|----|----|
|                           | Perfect                     | Imperfect                 |    |    |    |    |    |
| A                         | 6.43                        | 6.19                      | 6.81 | 5.23 | 5.37 | 4.39 | 4.57 |
| n                         | 4595                        | 2114                      | 3492 | 847  | 185  | 36  | 37  |
| df                        | 1                           |                           |     | 1   |     |     |     |
| H                         | 4.36                        |                           | 847 | 185  | 36   | 37  |     |
| P value                   | 0.0369                      |                           | 0.0245 | <0.0001 |     |     |     |
| H<sub>e</sub>             | 0.485                       | 0.486                     | 0.507 | 0.419 | 0.399 | 0.329 | 0.363 |<0.0001 |
| n                         | 4164                        | 1887                      | 3193 | 737  | 174  | 33  | 36  |
| df                        | 1                           |                           | 1    | 4   |     |     |     |
| H                         | 5.06                        |                           | 847 | 185  | 36   | 37  |     |
| P value                   | 0.8513                      |                           | 0.0245 | <0.0001 |     |     |     |
| Repeat frequency          |                             |                           | 15.20 | 9.65 | 8.85 | 5.31 | 6.47 |
| n                         | 4485                        |                            | 3427 | 810  | 176  | 36  | 36  |
| df                        |                            |                           | 1    | 4   |     |     |     |
| H                         |                            |                           | 847 | 185  | 36   | 37  |     |
| P value                   |                            |                           | 0.8513 | <0.0001 |     |     |     |
| Motif region              |                             |                           | 30.38 | 28.96 | 35.41 | 26.53 | 38.83 |<0.0001 |
| n                         | 4485                        |                            | 3427 | 810  | 176  | 36  | 36  |
| df                        |                            |                           | 1    | 4   |     |     |     |
| H                         |                            |                           | 847 | 185  | 36   | 37  |     |
| P value                   |                            |                           | 0.8513 | <0.0001 |     |     |     |
| Mean size<sup>b</sup>     |                             |                           | 201.25 | 204.50 | 198.03 | 208.00 | 225.45 | 207.79 | 216.03 |
| n                         | 4592                        | 2114                      | 3487 | 848  | 186  | 36  | 37  |
| df                        | 1                           |                           | 3427 | 810  | 176  | 36  |     |
| H                         | 3.91                        |                           | 847 | 185  | 36   | 37  |     |
| P value                   | 0.048                       |                           | 0.0084 | <0.0001 |     |     |     |

*Note: A = number of alleles; df = degrees of freedom; H<sub>e</sub> = expected heterozygosity; H<sub>o</sub> = observed heterozygosity.

*Pairwise comparisons between motif lengths are incorporated into Figs. 2 and 3 for different motif lengths.

*The statistics shown here are for the mean microsatellite size. The minimum and maximum microsatellite sizes show similar significant differences between motif types: Mean<sub>min</sub> Perfect = 184.3 bp, mean<sub>min</sub> Imperfect = 190.2 bp, H<sub>min</sub> = 6.54, P = 0.0105; mean<sub>max</sub> Perfect = 212.6 bp, mean<sub>max</sub> Imperfect = 220.0 bp; H<sub>max</sub> = 10.9, P = 0.0009.*
Even though microsatellite primers tested across genera or more distantly related species had been purposely removed from the data set, there were still 49 entries (24 microsatellite markers) for which the motifs and primer pairs matched across multiple loci. Thirty-eight entries were identified where matching primer sequences (forward and reverse) were found but with differing repeat motifs reported (19 primers total). In all of these instances, they were of the same genera. However, there was one primer with *Davidia involucrata* Baill. (Nyssaceae) and *Hedyotis chrysotricha* (Palib.) Merr. (Rubiaceae) where the primer sequences matched, although they did have differing repeat motifs. These matches among primer pairs do not take into account alternative motif permutations.

**Taxonomy**—There was no significant difference between monocots and dicots in any of the measures of genetic variation (A, *H*<sub>e</sub>, and *H*<sub>o</sub>). The monocots did, however, have significantly larger motif regions compared to dicots (*H* = 17.3, *P* < 0.0001; see Table 3). Across the different plant taxonomic clades, the gymnosperms exhibited a significantly greater number of alleles than most other plant taxonomic clades, whereas the eudicots, asterids, and rosids had significantly reduced heterozygosity and number of alleles than most of their evolutionarily older counterparts, with the exception of the Nymphaeales (Table 3, Fig. 6).

**DISCUSSION**

The primary goal of this project was to identify specific characteristics of microsatellites that may aid researchers in choosing effective markers for applications requiring genetic
variation, such as quantifying population genetic structure and diversity, estimates of mating systems, and paternity analysis. Now that library development and the isolation of putative microsatellite markers have become relatively straightforward, the remaining challenge in the development process is choosing which markers to further investigate and screen for amplification success and polymorphism. Researchers could better focus their time and effort if they knew specific characteristics of microsatellite markers that are associated with higher levels of genetic variation. Here we generate and use a data set for an empirical review on microsatellite markers that have been developed over the past 18 years, to identify relationships across higher taxa, and conclude with specific recommendations for marker selection.

Markers containing dinucleotide repeats exhibited significantly higher levels of genetic variation in $A$, $H_e$, and $H_o$ than most other motif lengths (Table 2, Figs. 2 and 3). This is consistent with other studies that suggest dinucleotide repeats are generally more variable than other motif lengths, most likely due to the relative ease of mutation via DNA slippage during replication (Levinson and Gutman, 1987; Grist et al., 1993; Strand et al., 1993; Tautz and Schlötterer, 1994; Chakraborty et al., 1997; Sup Lee et al., 1999; Ellegren, 2000, 2004). This slippage is also a potential disadvantage to dinucleotide repeats as it can lead to difficulty in scoring alleles on an electropherogram (i.e., more stutter peaks) compared to larger motif lengths (e.g., Brown et al., 1996). In addition, the differences found here may also be due in part to the very large number of dinucleotide microsatellites reported in the literature relative to all other motif lengths (approximately 73%). Although the overrepresentation of dinucleotides can potentially bias the statistical analysis, the Kruskal–Wallis test incorporates sample size and, therefore, unequal sample sizes should be of minimal concern, especially considering the large number of samples within each motif length (e.g., mean number of entries of all motif lengths = 895). Furthermore, even if investigators have been influenced over the years toward choosing dinucleotide repeat motifs, their popularity may be due in large part to the natural prevalence of dinucleotide repeats throughout plant genome (Brown et al., 1996; Zhao and Ganal, 1996; but see Morgante et al., 2002).

### Table 3. Group comparisons of major taxonomic clades. Each row includes means and number of entries ($n$) for each grouping along with Kruskal–Wallis statistics for each comparison made ($H$ and the corresponding $P$ value), including cotyledon type and major taxonomic rankings. Significant ($P < 0.05$) values are shown in bold.

| Variable          | Cotyledons | Taxonomic clades<sup>a</sup> |
|-------------------|------------|-----------------------------|
|                   | Dicots     | Monocots                    | Gymnosperms | Nymphaeales | Austrobaileyales | Magnoliids | Eudicots | Rosids | Asterids |
| $A$ Mean          | 6.33       | 6.49                        | 7.75        | 4.97        | 14.78           | 5.68       | 5.81     | 6.21   | 6.63     |
| $n$               | 4374       | 1158                        | 416         | 31          | 9               | 270        | 307      | 2447   | 1615     |
| $df$              | 1          | 1                           | 1           | 1           | 1               | 1          | 1        | 1      | 1        |
| $H$ Mean          | 0.0533     | 0.8175                      | 0.8113      | 0.425       | 0.813           | 0.485      | 0.570    | 0.579   | 0.564    |
| $P$ value         |            |                             | 0.8175      | 0.425       | 0.813           | 0.485      | 0.570    | 0.579   | 0.564    |
| $H_e$ Mean        | 0.573      | 0.588                       | 0.619       | 0.425       | 0.813           | 0.485      | 0.570    | 0.579   | 0.564    |
| $n$               | 4194       | 1077                        | 412         | 29          | 9               | 270        | 280      | 2402   | 1512     |
| $df$              | 1          | 1                           | 1           | 1           | 1               | 1          | 1        | 1      | 1        |
| $H$ Mean          | 0.968      | 0.968                       | 0.968       | 0.968       | 0.968           | 0.968      | 0.968    | 0.968   | 0.968    |
| $P$ value         | 0.3251     |                             | 0.3251      | 0.3251      | 0.3251          | 0.3251     | 0.3251   | 0.3251  | 0.3251   |
| Repeat frequency  |            |                             |            |             |                 |            |          |        |          |
| $A$ Mean          | 13.5       | 13.8                        | 16.0        | 20.2        | 23.8            | 13.7       | 13.8     | 13.5   | 13.5     |
| $n$               | 2860       | 857                         | 304         | 13          | 8               | 150        | 185      | 1731   | 942      |
| $df$              | 1          | 1                           | 1           | 1           | 1               | 1          | 1        | 1      | 1        |
| $H$ Mean          | 1.86       |                             | 1.86        | 1.86        | 1.86            | 1.86       | 1.86     | 1.86   | 1.86     |
| $P$ value         | 0.1770     |                             | 0.1770      | 0.1770      | 0.1770          | 0.1770     | 0.1770   | 0.1770 | 0.1770   |
| Motif region      |            |                             |            |             |                 |            |          |        |          |
| $A$ Mean          | 29.3       | 31.0                        | 34.7        | 40.3        | 47.5            | 29.3       | 29.6     | 29.2   | 29.4     |
| $n$               | 2860       | 857                         | 304         | 13          | 8               | 150        | 185      | 1731   | 942      |
| $df$              | 1          | 1                           | 1           | 1           | 1               | 1          | 1        | 1      | 1        |
| $H$ Mean          | 17.3       |                             | 17.3        | 17.3        | 17.3            | 17.3       | 17.3     | 17.3   | 17.3     |
| $P$ value         | <0.0001    |                             | <0.0001     | <0.0001     | <0.0001         | <0.0001    | <0.0001  | <0.0001| <0.0001  |
| Mean size         |            |                             |            |             |                 |            |          |        |          |
| $A$ Mean          | 198.8      | 206.2                       | 214.1       | 165.0       | 207.1           | 192.5      | 196.2    | 197.0   | 201.9    |
| $n$               | 4379       | 1157                        | 426         | 31          | 9               | 270        | 307      | 2451   | 1616     |
| $df$              | 1          | 1                           | 1           | 1           | 1               | 1          | 1        | 1      | 1        |
| $H$ Mean          | 1.504      |                             | 1.504       | 1.504       | 1.504           | 1.504      | 1.504    | 1.504  | 1.504    |
| $P$ value         | 0.2201     |                             | 0.2201      | 0.2201      | 0.2201          | 0.2201     | 0.2201   | 0.2201 | 0.2201   |

Note: $A$ = number of alleles; $df$ = degrees of freedom; $H_e$ = expected heterozygosity; $H_o$ = observed heterozygosity.

<sup>a</sup>Statistics comparing taxonomic clades include the monocots; however, the means for this group are listed under the Cotyledons column heading.
To identify traits conferring greater levels of genetic variation with certain microsatellite markers, primer pairs generating monomorphic loci were included as a baseline for the number of alleles and heterozygosity values. In total, there were only 532 (7.67%) reports of monomorphic markers in the database; this is likely an underrepresentation of the actual proportion in nature because many authors and some journals prefer to exclude monomorphic primers from publication, which biases the data set. Given the importance of monomorphic markers in this and other studies, we recommend that researchers report data from these markers whenever possible. When isolating the more common number of alleles (one to seven), we found that monomorphic markers were more likely to be associated with dinucleotide motif lengths (65.4% of monomorphic primers were dinucleotide) than other motif lengths. However, when put into perspective, the vast majority of microsatellite markers are dinucleotide repeats (72.5%; Fisher’s exact, $P < 0.0001$), and of these, only 8.67% are monomorphic, compared to 9.6–19.0% of markers possessing three to seven alleles. In addition, some cases of monomorphism may indicate the presence of unseen null alleles, which unfortunately are not always reported in the literature and may be difficult to detect accurately; therefore, null alleles could not be included in our analysis.

Many studies have identified the most abundant repeat motifs reported in the literature (Morgante and Olivieri, 1993; Wang et al., 1994; Brown et al., 1996; Zhao and Ganal, 1996; Zane et al., 2002). Contrary to the suggestion that AT repeats should be preferred (Morgante and Olivieri, 1993), here we found that the dinucleotide motif $GA_n$ was the most abundant, both as a unique motif and including (in descending order of frequency) the complement ($CT_n$), reverse ($AG_n$), and reverse complement ($TC_n$). This was similar to previous studies in which GA (or AG) is often reported as one of the most abundant repeat motifs (e.g., Wang et al., 1994; Zane et al., 2002). Furthermore, $CTT_n$ was the most abundant trinucleotide repeat, along with the reverse complement ($AAG_n$), complement ($GAA_n$), and reverse ($TTC_n$).

Significantly higher levels of genetic variation ($A$ and $H_e$) were found in perfect motif types compared with imperfect motif types (Table 2). This finding corroborates previous suggestions that interrupted motifs reduce stutter and therefore result from mechanisms of mutation (e.g., due to slippage in replication; Richards and Sutherland, 1992, 1994; Pépin et al., 1995; Petes et al., 1997; Rossetto, 2001). It should be noted that the inclusion of compound and interrupted repeats into one (imperfect) category may have skewed the levels of genetic variation detected here for that group of motif types; further study isolating these more-specific motif types may ameliorate the effects of having combined these within a single group. In our study, we were most interested in characteristics of perfect microsatellites; future investigations may wish to focus on subtle variations between compound and interrupted motifs.
The strong correlations between $A$, $H_s$, and $H_o$ in the data set were to be expected, given that they all are measures of genetic diversity. The reduced correlation between $A$ and $H_o$ ($r_s = 0.530$) compared to $A$ and $H_s$ ($r_s = 0.835$) may be attributed to $H_o$ serving as a direct trait specific to the populations tested as opposed to being a trait of the marker used. The negative correlation between microsatellite motif length with $A$, $H_s$, and $H_o$ strongly supports the suggestion that polymorphism in microsatellite markers is generally found in shorter motif sequences, which are more likely to slip (Grist et al., 1993; Chakraborty et al., 1997; Sup Lee et al., 1999; Ellegren, 2004). The positive correlations detected between repeat frequency or motif region with $A$, $H_s$, and $H_o$ suggest that longer repetitive sequences tend to result in greater polymorphism, regardless of mutation mechanisms (Weber and May, 1989; Valdes et al., 1993; Primmer et al., 1996; Ellegren, 2004). One might expect that microsatellite size range and motif length would be dependent upon one another. However, the lack of a relationship detected here suggests that the associated flanking regions around the motif may play a more important role in determining the size of the overall microsatellite fragment. It should also be noted that microsatellite size incorporates the forward and reverse primer lengths, which are selected by the researcher during primer development. Preliminary analysis, removing respective primer lengths for each microsatellite size range, showed no relationship of primer length with microsatellite size or with levels of heterozygosity. Therefore, primer lengths were included in microsatellite size for the remainder of the study. Although the magnitude of correlations between minimum and maximum microsatellite sizes with $A$, $H_s$, and $H_o$ are low, the directionality (inversely with minimum size and positively maximum size) suggests that greater overall lengths of microsatellite sizes will result in greater levels of genetic variability (e.g., a small minimum size and a large maximum size). This is consistent with the expectation that as the size of the marker increases, so too does the repeat frequency and therefore the number of possible alleles. However, the low magnitude of these correlations suggests that there is a point at which this relationship breaks down.

In this study, we intentionally removed genic (EST) SSRs to focus on genomic markers. While genic markers provide distinct advantages over genomic markers, including cross-species compatibility due to conservation of transcribed regions and a high generation rate at low costs through mining data repositories (Varshney et al., 2005), genomic markers typically provide greater levels of genetic variation (Cho et al., 2000; Scott et al., 2000; Eujayl et al., 2001; Rungis et al., 2004; Russell et al., 2004; Chabane et al., 2005; Woodhead et al., 2005; Martin et al., 2010), present fewer null alleles (Varshney et al., 2005), and are less likely to be subject to direct or artificial selection (in the case of agricultural crops). This is not to say that genomic markers are not without their own drawbacks, including intensive time and resource commitment in isolation, relatively reduced cross-species compatibility, and the inability to inform phenotypic expression. Future investigations could focus on genic markers and examine how their characteristics are associated with measures of genetic variation. This would be helpful as previous studies have suggested that using both genic and genomic markers in concert is a more powerful approach than choosing one marker type over another (e.g., Martin et al., 2010).

Given that our data set included marker and trait information across species and families, we wished to exercise its value by examining variation across higher taxonomic groups. Although there are inconsistent patterns in the variation of microsatellite markers across taxonomic clades, more general trends can be described here. First, there was an underrepresentation of gymnosperms and some extant angiosperms (Nymphaeales and Austrobaileyales) in the literature of reported microsatellite markers. The combined mean number of entries (e.g., microsatellite markers) for the gymnosperms and basal angiosperms was approximately 151, compared to more than 1000 in the angiosperms (monocots, true eudicots, rosids, and asterids), including the magnoloids. This is in part due to the limited number of species (and in some cases, even individuals within species) in these more extant clades, but it also highlights an overall popular interest in angiosperms. Second, the higher genetic variation found in gymnosperms compared to most angiosperm clades may in one sense be attributed to their greater evolutionary age. However, as much as SSRs are thought to form via mutation by expansion due to slippage, there are just as easily contractions of tandemly repeated fragments that reduce the size of SSRs, or gradual elimination of repetitive sequences due to point mutations, given enough time (e.g., into cryptically simple motifs; Tautz et al., 1986; Hancock, 1999). In addition, this variability between gymnosperms and angiosperms may be due to greater genetic admixture among species, especially considering the typical wind-pollination strategies of the gymnosperm clade (Faegri and van der Pijl, 1979) compared to the smaller subset of angiosperms that are wind-pollinated. Although the high genetic diversity reported in Austrobaileyales may reflect its taxonomic age, it is more likely due to a sampling bias as there were only nine markers available for the single species in this clade. In the cases of other taxonomic clades, there may be more subtle, species-specific mechanisms at play. For instance, it has been found in eukaryotes (Chang et al., 2001) and bacteria (LeClerc et al., 1996) that mutations within mismatch repair genes may provide an evolutionary benefit to increased or decreased mutability in gene sets, depending on conferred advantages or disadvantages with such mutations. Species-specific mismatch repair genes in different plant taxa may likewise play a role in the widespread lack of consistent patterns across taxonomic clades.

**Related discoveries**—Of the 6782 microsatellite markers in the database, there was only one that matched primer pairs (but the motif differed) across different genera and even families, without considering reverse, complementarity, and reverse-complementarity of reported markers. This was between *Davidia involucrata* (Nyssaceae) and *Hedyotis chrysotricha* (Rubiaceae). This match is in part due to the primer pair coming from the same laboratory (Du et al., 2012; Yuan et al., 2012), but its rarity also illustrates the limited success of cross-species amplification with genomic markers (Rubinsztein et al., 1995; Primmer et al., 1996; Morin et al., 1998) as compared with genic markers (e.g., Varshney et al., 2005). Ellegren et al. (1995) suggest that while genomic loci may be orthologous among related species, the polymorphism about those loci are not generally conserved as evolutionary age increases.

In this study, we did not differentiate between compound and interrupted motifs because we were more interested in finding characteristics of perfect motifs that may aid investigators in choosing which markers to further pursue during the de novo development phase. However, it is worth noting that interrupted motifs have been suggested as a better choice of marker when comparing more distantly related species as they do show a tendency to mutate at a slower rate compared to more instable...
tandem repeats of perfect motifs (Rossetto, 2001). Further examination of the characteristics of the various types of imperfect motifs in relation to measures of genetic diversity would be informative.

There was another small number of entries where primer pairs matched—with or without matching motifs—that did not warrant outright removal in the initial steps of compiling the database. All of these were within-species duplicates and occurred in multiple loci. In cases where the motif differed in composition (38 entries total, 19 primer pairs), seven motifs differed in repeat frequency, eight contained entirely different motif configurations, one was a complementary motif of equal length, and the last case consisted of three complementary motifs with differing repeat frequencies. The chances of obtaining such identical primer sequences are very low. For example, forward and reverse primer regions containing the repeated sequence of interest are usually chosen to be 15–30 bp in length by the investigator. The minimum odds of finding this same motif confi guration may be the mutagenic force behind these particular markers. NGS technology may be useful in further characterizing the role of recombination and crossing over in microsatellite evolution through identifying like microsatellite markers across multiple loci in genomes.

Conclusions—In this study, we compiled a large, publicly available data set of characteristics of microsatellite markers published over the past 18 years and showed how these traits are associated with levels of genetic variation. This information can now be used to aid researchers developing new microsatellite markers to conserve their time and resources by choosing the most effective markers for population screening. We also used the data set in a preliminary study to identify trends in levels of genetic variation across major taxonomic groups. While this was only an initial analysis, we encourage further research using this data set to explore levels of genetic variability within and across specific taxonomic families. Other potential uses of the database could include looking for associations between motif lengths and null alleles, or examining differences in markers that are in Hardy–Weinberg equilibrium vs. those that are not; both would necessitate revisiting the original literature to quantify null allele presence or Hardy–Weinberg conformity. In addition, the current study focused on genomic markers, but the workflow described here could also be used to examine analogous characteristics of EST-SSRs. Considering the myriad of EST-SSR resources now available via online databases for nonmodel systems, analyses of EST data would require less time-intensive manual data entry than that described here for genomic markers. Finally, other potential uses of the data set include further exploration of theories into the evolution of repetitive DNA.

Although researchers may benefit from including a variety of different types of microsatellite markers in their genetic investigations, several general conclusions can be drawn from the empirical evidence presented here and in the literature. More generally, attention should be given to comparing genetic variation across studies using different motif lengths, as conclusions may vary due in part to the characteristics of the microsatellites rather than only the natural variation present. We recommend that researchers developing primers for fine-grain analysis of population genetic structure and analysis or mating system estimates should focus on dinucleotide repeats exhibiting a large repeat frequency and wide-ranging overall microsatellite fragment size. When working on a more coarse scale with more distantly diverged (either geographically or temporally) species or taxa, the use of either interrupted repeats or lower repeat frequency perfect motifs may aid in capturing the slower mode and tempo of change while still retaining some degree of relatedness. Microsatellites continue to be important and relevant in a wide variety of studies. Therefore, we recommend that researchers carefully consider the characteristics of the markers that they choose to develop with respect to the types of studies they are intended for, rather than randomly selecting primer pairs to further test in the microsatellite development process.

LITERATURE CITED

Abdelkrim, J., B. Robertson, J.-A. Stanton, and N. Gemmell. 2009. Fast, cost-effective development of species-specific microsatellite markers by genomic sequencing. BioTechniques 46: 185–192.

Bhargava, A., and F. F. Fuentes. 2010. Mutational dynamics of microsatellites. Molecular Biotechnology 44: 250–266.

Brown, S. M., A. K. Szewc-McFadden, and S. Kresovich. 1996. Development and application of simple sequence repeat (SSR) loci for plant genome analysis. In P. F. Jahar [ed.], Methods of genome analysis in plants, 147–159. CRC Press, New York, New York, USA.

Buschiazzo, E., and N. J. Gemmell. 2006. The rise, fall and Renaissance of microsatellites in eukaryotic genomes. BioEssays 28: 1040–1050.

Chabane, K., G. A. Ablett, G. M. Cordeiro, J. Valkoun, and R. J. Henry. 2005. EST versus genomic derived microsatellite markers for genotyping wild and cultivated barley. Genetic Resources and Crop Evolution 52: 903–909.

Chakkaborty, R., M. Kimmel, D. N. Stivers, J. D. Davison, and R. DeK. 1997. Relative mutation rates at di-, tri-, and tetranucleotide microsatellite loci. Proceedings of the National Academy of Sciences, USA 94: 1041–1046.

Chang, D. K., D. Metzgar, C. Wills, and C. Bolland. 2001. Microsatellites in the eukaryotic DNA mismatch repair genes as modulators of evolutionary mutation rate. Genome Research 11: 1145–1146.

Cho, Y. G., T. Ishi, S. Temnykh, X. Chen, L. Lipovich, S. R. McCouch, W. D. Park, et al. 2000. Diversity of microsatellites derived from genomic libraries and GenBank sequences in rice (Oryza sativa L.). Theoretical and Applied Genetics 100: 713–722.

Du, Y.-J., Q.-Y. Dai, L.-Y. Zhang, Y.-X. Qiu, J.-H. Li, and C.-X. Fu. 2012. Development of microsatellite markers for the dove tree, Davidia involucrata (Nyssaceae), a rare endemic from China. American Journal of Botany 99: e206–e209.

Ekblom, R., and J. Galindo. 2011. Applications of next generation sequencing in molecular ecology of non-model organisms. Heredity 107: 1–15.

Ellegren, H. 2000. Heterogeneous mutation processes in human microsatellite DNA sequences. Nature Genetics 24: 400–402.

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

Ellegren, H., C. R. Primmer, and B. Sheldon. 1995. Microsatellite evolution: Directionality or bias in locus selection? Nature Genetics 11: 360–362.

Estoup, A., and B. Angers. 1998. Microsatellites and minisatellites for molecular ecology: Theoretical and empirical considerations. In G. R. Carvalho [ed.], Advances in molecular ecology, 55–85. IOS Press, Burke, Virginia, USA.

Euiyai, I., M. Sorrells, M. Baum, P. Wolters, and W. Powell. 2001. Assessment of genotypic variation among cultivated durum wheat based on EST-SSRS and genomic SSRs. Euphytica 119: 39–43.

Fager, K., and L. van der Pul. 1979. The principles of pollination ecology, 3rd ed. Pergamon Press, Oxford, United Kingdom.

Goldstein, D. B., and C. Schlotterer. 1999. Microsatellites: Evolution and applications. Oxford University Press, Oxford, United Kingdom.

Griff, S. A., F. A. Figuiera, and A. A. Morley. 1993. Dinucleotide repeat polymorphisms isolated by the polymerase chain reaction. BioTechniques 15: 304–309.
using AFLPs and SSRs from anonymous and transcribed gene regions. *Molecular Ecology* 14: 1681–1695.

Yuan, N., H. P. Comes, Y.-R. Mao, X.-S. Qi, and Y.-X. Qiu. 2012. Genetic effects of recent habitat fragmentation in the Thousand-Island Lake region of southeast China on the distyloous herb *Hedyotis chrysotricha* (Rubiaceae). *American Journal of Botany* 99: 1715–1725.

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

Zhao, X., and M. W. Ganal. 1996. Applications of repetitive DNA sequences in plant genome analysis. In Andrew H. Paterson [ed.], *Genome mapping in plants*, 111–125. Academic Press, San Diego, California, USA.

**APPENDIX 1. Glossary of terms.**

| Term                        | Definition                                                                                                                                 |
|-----------------------------|-------------------------------------------------------------------------------------------------------------------------------------------|
| **Flanking sequence**       | The nucleotides found immediately on either side of the repeated motif within the microsatellite fragment.                                  |
| *ACGTG*TATATATATATATATA*GAGG | (forward primer)ACGTGTTATATATATATATATA*GAGG(reverse primer)                                                                                 |
| **Microsatellite marker**   | The entire sequence containing the forward and reverse primers, the repeated motif of interest, and any intervening flanking sequences. This does not include fluorescent markers used to identify fragments. |
| *ACGTG*TTATATATATATATATA*GAGG | (forward primer)ACGTGTTATATATATATATATA*GAGG(reverse primer)                                                                                 |
| **Motif**                   | Nucleotide composition of the repeated sequence.                                                                                           |
| *ACGTG*TATATATATATATATA*GAGG | (forward primer)ACGTGTTATATATATATATATA*GAGG(reverse primer)                                                                                 |
| **Motif length**            | The number of bases within the repeated motif.                                                                                              |
| *di-* TA*<sub>n</sub>, tri-* CGG*<sub>n</sub>, tetra-* GAAT*<sub>n</sub>, etc.* |                                                                                                                                             |
| **Motif type**              | The arrangement of the repeated motif; these can be perfect [(CA)<sub>n</sub> or (GTAG)<sub>n</sub>], compound [(AT)<sub>n</sub>(GTC)<sub>n</sub>], or interrupted [(TC)<sub>n</sub>CTAG(CCG)<sub>n</sub>]. For our study, compound and interrupted repeat types are known as imperfect. |
| **Motif region**            | The size range of the repeated motif.                                                                                                        |
| ACGTGTTATATATATATATATA*GAGG | (TA)<sub>n</sub> = 2 bp x 9 = 18                                                                                                               |
| **Repeat frequency**        | The number of times (n) that a motif is repeated within a fragment.                                                                          |
| *TA*<sub>n</sub>            |                                                                                                                                              |