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
Department of Biological Sciences, University of Cincinnati

Introduction

Microsatellite markers (e.g. simple sequence repeats; SSRs) continue to be the marker of choice in various studies addressing genetic diversity and structure, paternity analysis, and mating system estimates. Despite the benefits of these markers, a considerable drawback is that generally markers must be developed *de novo* for a species, even if markers are available in closely related species. The several hundreds of putative markers isolated through traditional library development and the thousands that can be identified in next generation sequencing need to be further tested for amplification success in the species of interest. How best should a researcher proceed in choosing from these massive numbers of potential markers? While there is no currently accepted criteria for marker selection, we compiled a large database of microsatellite markers developed over the larger part of the last twenty years to empirically identify traits conferring higher levels of genetic diversity. With this dataset, we sought to answer 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?

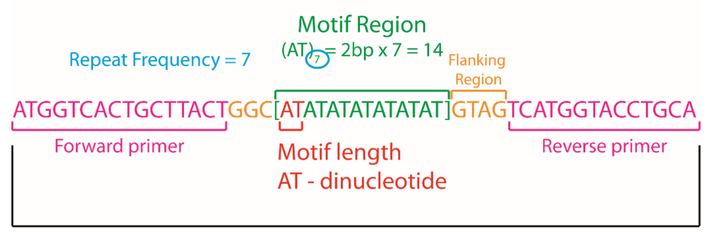
Abstract

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 6,782 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 fragment size. Variation across taxonomic lineages 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. A preliminary study using this dataset, looking at relationships in genetic variability across higher taxa showed that higher taxonomical lineages were disproportionately represented in the literature and exhibited little consistency.

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.

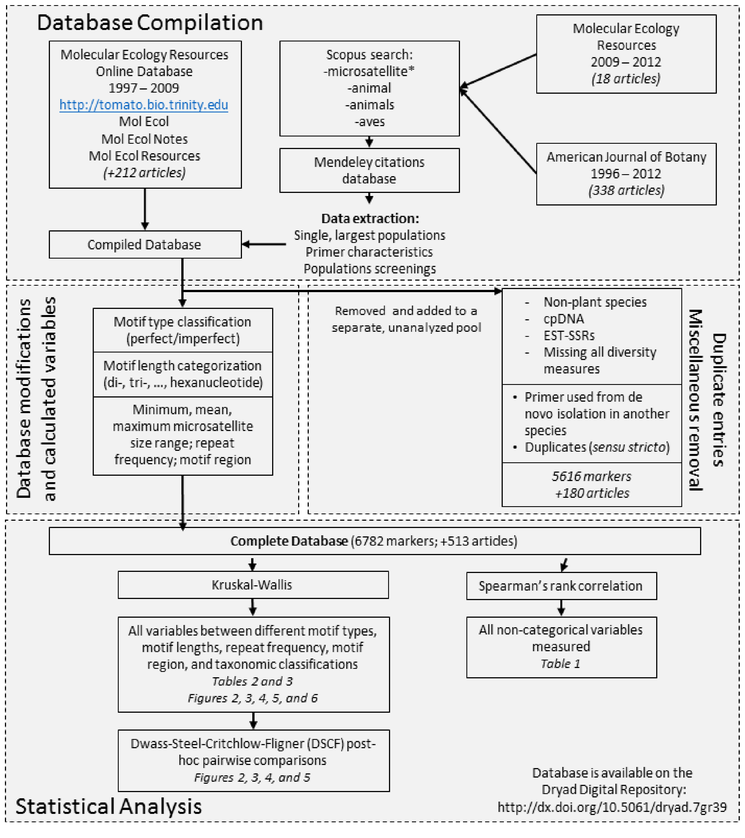
Microsatellite marker characteristics

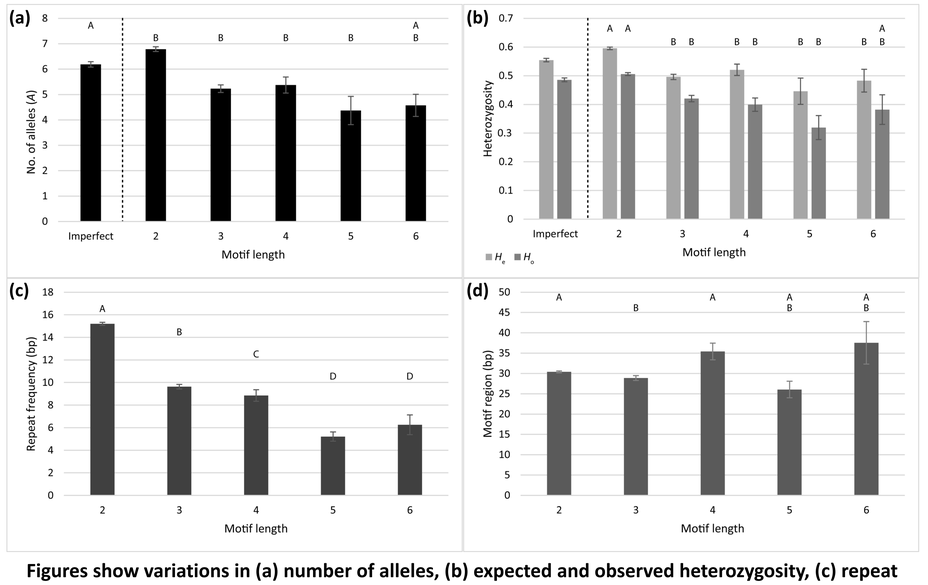


Motif type – the arrangement of the repeated motif; these can be perfect [$(CA)_n$ or $(GTAG)_n$], compound [$(AT)_n(GTC)_n$], or interrupted [$(TC)_nCT(CCG)_n$]. Here, compound and interrupted repeat types are known as imperfect.
Motif length – the size of the repeated fragment [e.g. $(AT)_n$ = dinucleotide, $(GTC)_n$ = trinucleotide]
Repeat frequency – the number of time the motif length is repeated [e.g. the *7* in $(AT)_7$ or the *n* in $(GTC)_n$]
Motif region – the overall length of the repeated segment [e.g. $(AT)_7$ = 2bp x 7 = 14]
Flanking Region – The nucleotides found immediately on either side of the repeated motif region within the microsatellite fragment.
Microsatellite size – the overall length of the microsatellite fragment, including forward and reverse primers, any intervening flanking regions, and the repeated motif of interest.

Methods

The dataset was compiled from an online database and manually mined from hundreds of published articles, resulting in a total database of over 6000 unique primers from more than 500 primer notes and articles.





Figures show variations in (a) number of alleles, (b) expected and observed heterozygosity, (c) repeat frequency, and (d) motif region length across different motif lengths. Letters above figures indicate Dwass-Steel-Critchlow-Fligner post-hoc comparison groupings (a non-parametric equivalent of Tukey's HSD). In (a) and (b) imperfect motifs are included for comparison but were excluded from statistical tests comparing motif lengths.

Results

- Compared with imperfect motifs, perfect motifs exhibited significantly higher levels of A and H_e (Kruskal-Wallis H = 4.36 and 5.06; P = 0.037 and 0.025, respectively).
- Within perfect motifs, motif lengths differed significantly from one another for A, H_e, and H_o (H = 107.89, 132.96, and 82.08; P <0.0001, respectively). The dinucleotide repeat motifs exhibited significantly higher H_e than any other motif length, and significantly higher A and H_o than the tri-, tetra- and pentanucleotide repeats (see figures below).
- Repeat frequency exhibited positive significant correlations with A, H_e, and H_o (Spearman's rank correlation r_s = 0.413, 0.395, 0.246; P <0.0001; respectively; see figure). Motif region differed significantly across motif lengths (H = 28.4, P <0.0001) however there was no consistent trend (see figure).
- Microsatellite size significantly differed across motif lengths (H = 39.6, P<0.0001) with a general trend of size increasing (data available in publication). However, there was no correlation between mean microsatellite size and A, H_e, or H_o.
- 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) are often reported as one of the most abundant repeat motifs (e.g. Wang et al., 1994 or Zane et al., 2001).

Discussion

Significantly higher levels of genetic variation (A and H_e) were found in perfect motif types compared with imperfect motif types. This finding corroborates previous suggestions that interrupted motifs reduce stutter and therefore result from mechanisms of mutation (e.g. due to slippage in replication; e.g. Rossetto, 2001). Our finding that dinucleotide repeats exhibit significantly higher levels of genetic variation compared with most other motif types 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 (e.g. Levinson and Gutman, 1987; Ellegren, 2004). The positive associations between repeat frequency and motif region with levels of genetic variation (A, H_e, and H_o) suggest that researchers should focus more on large, highly repetitive fragments. Depending on the type of study researchers wish to conduct using SSRs, the lack of variability in markers may play to their benefit, where lower levels of variability in markers can capture the slower mode and tempo of mutation across taxa (e.g. Rossetto, 2001), as compared to higher levels of variability with dinucleotide repeats exhibiting a high repeat frequency that may be helpful in differentiating population genetic structure or parentage analysis. Therefore, we suggest that researchers carefully consider the type of study they wish to employ SSRs in and use those criteria to guide marker selection. Please see our publication this August in Applications in Plant Sciences for a more detailed synthesis and analysis of the dataset, and take advantage of the dataset available on the Dryad digital repository: http://dx.doi.org/10.5061/dryad.7gr39.

References:

Ellegren, H. 2004. Microsatellites: simple sequences with complex evolution. *Nature Reviews. Genetics* 5: 435–445.
Levinson, G., and G. A. Gutman. 1987. Slipped-strand mispairing: a major mechanism for DNA sequences evolution. *Molecular Biology and Evolution* 4: 203-221.
Rossetto, M. 2001. Sourcing of SSR markers from related plant species. *In* R. Henry [ed.], Plant genotyping: the DNA fingerprinting of plants, 211-224. CABI, Wallingford, UK.
Zane, L., L. Bargelloni, and T. Patarnello. 2002. Strategies for microsatellite isolation: a review. *Molecular Ecology* 11: 1–16.
Wang, Z., J. L. Weber, G. Zhong, and S. D. Tanksley. 1994. Survey of plant short tandem DNA repeats. *Theoretical and Applied Genetics* 88: 1-6.