# Insertions-Deletions in a Microsatellite Flanking Region May Be Resolved by Variation in Stuttering Patterns 

Felix Gugerli • Sabine Brodbeck • Rolf Holderegger

Published online: 17 May 2008
(C) Springer-Verlag 2008


#### Abstract

Microsatellites or simple sequence repeats (SSRs) may display polymerase-chain-reaction-amplified fragment lengths mismatching the patterns expected from repeat copy number variation. We sequenced alleles of a nuclear dinucleotide SSR locus in two oak species which showed 2- and 1-bp length differences between alleles and three types of stuttering patterns in fragment length analysis. In accordance with the variation in stuttering, we identified three allele classes characterized by insertions-deletions in the flanking regions and overlapping repeat copy number ranges. Different alleles could thus only be safely separated when considering these stuttering patterns. Our results raise the question of how to adequately delimit alleles when such size homoplasy is present. We advise to thoroughly characterize SSR sequence variation during marker development and to carefully place primer sites along flanking regions to facilitate automated allele scoring and to minimize laborintensive visual inspection.


Keywords Allele calling • Flanking region • Fragment length analysis •
Nuclear microsatellite • Nucleotide substitutions • Simple sequence repeat .
Size homoplasy • Stuttering pattern

## Abbreviations

bp base pair
FLA fragment length analysis
indel insertion-deletion
nSSR nuclear simple sequence repeat
PCR polymerase chain reaction

[^0]
## Introduction

Microsatellites or simple sequence repeats (SSRs) are numerous and spread across the genomes of eukaryotic organisms, which makes them a useful tool for many types of population genetic studies (Zhang and Hewitt 2003; Selkoe and Toonen 2006) or for genome mapping (Schlötterer 1998) in plants and animals. Owing to their great variability, nuclear SSRs (nSSRs) have proven particularly informative when studies require codominance and high resolution, e.g., kinship structure, parentage analysis, population assignment, or genetic fingerprinting (Goldstein and Schlötterer 1999).

Homoplasy is a problem when applying SSRs to population genetic studies because alleles considered identical in state are not necessarily identical by descent (Estoup et al. 2002). Moreover, variation in flanking regions may lead to size homoplasy beyond that imposed by mutations within the microsatellite region. This effect occurs when insertions-deletions (indels) in flanking regions compensate for differences in repeat copy numbers among different alleles. Likewise, indels that deviate in length from the multiple of the repeat motif result in allele sizes that do not fit into the pattern of regular allele size differences. Careful evaluation of fragment lengths during scoring and allele binning is therefore mandatory, particularly in studies that require accurate kinship identification (e.g., pedigree, parentage, forensics). Though these problems and their potential consequences for biological interpretation have long been recognized (Hoffman and Amos 2005), there is only limited information on the nature of sequence variation in SSR-comprising fragments and on how to circumvent allele misidentification. Hence, most population genetic studies using nSSRs ignore respective pitfalls (Selkoe and Toonen 2006).

In the course of population genetic analyses of oaks (Quercus petraea (Matt.) Liebl., Quercus robur L.; Gugerli et al. 2007), the nSSR locus ssrQrZAG30 (Kampfer et al. 1998) revealed fragment lengths among alleles that partly deviated from those expected in a dinucleotide repeat. Within the total allele size range of $171-257 \mathrm{bp}$, we found regular 2-bp differences between distinct allelic bins up to 201 bp . However, in allelic bins longer than 201 bp, we observed almost gradual length increases. We further noticed that allelic bins differing by about 1 bp showed stuttering patterns that could be distinguished owing to the varying number and intensity of stutter bands (Fig. 1). These stuttering patterns could be replicated across separate polymerase chain reaction (PCR) amplifications and in offspring tests, which indicated a sequence-induced cause of different stuttering patterns and their Mendelian inheritance (data not shown). We therefore sequenced alleles spanning the entire allele size range (1) to disentangle the cause and nature of the unexpected pattern in fragment lengths and (2) to allow for correct allele calling taking into account variation in corresponding stuttering patterns.

## Materials and Methods

We originally PCR-amplified the nSSR locus ssrQrZAG30 with a fluorescently labeled primer, using genomic DNA extracted from leaf tissue collected in a mixed oak stand in Switzerland (Gugerli et al. 2007). Fragment length analysis (FLA) on an ABI3100-Avant capillary sequencer (Applied Biosystems) relied on an internal size

[^1]

Fig. 1 Fragment lengths (in assending order) and stuttering patterns observed at the microsatellite locus ssrQrZAG30 (Kampfer et al. 1998) in oaks (Quercus spp.). Stuttering patterns of the small allele class (black diamonds) are shown below data points, those of the longer two allele classes (medium class: gray diamonds; large class: open circle) above the data points
standard (ROX400HD; Applied Biosystems) using GeneScan 3.1 (Applied Biosystems) to score fragment lengths. We delimited allelic bins on the basis of a graph of all amplified fragments sorted according to their lengths (Fig. 1).

A first batch of ten alleles was cloned (pcr4-TOPO; Invitrogen) and sequenced by a commercial provider (ecogenics, Zürich-Schlieren). Another six alleles were sequenced directly from PCR amplification products of homozygous individuals or, in heterozygotes, after separate band extraction from agarose gels (MinElute Gel Extraction Kit, Qiagen). The remaining 24 alleles were cloned using a pGEM-T Easy Vector System (Promega) and sequenced. Before sequencing positive clones, we performed PCRs using the original nSSR primers to determine the fragment lengths of the respective inserts by FLA on the automatic sequencer. This allowed us to determine whether the selected positive clone showed the real fragment or rather a stutter fragment (see below). Subsequently, one of the positive clones with the longest fragment length for a given allele was further propagated overnight and extracted using a Wizard Plus SV Minipreps DNA-Purification System (Promega) and sequenced. Cycle sequencing was performed in $10-\mu \mathrm{L}$ reactions containing $2 \mu \mathrm{~L}$ BigDye Terminator V3.1 and $2 \mu \mathrm{~L} 5 \times$ dilution buffer (Applied Biosystems) on a PTC-100 thermocycler (MJ Research) using the cycle profile recommended by the manufacturer. Sequences were determined on an ABI3100Avant (Applied Biosystems) using POP4. We obtained double-strand consensus sequences of each allele from forward and reverse primers using AutoAssembler 1.4.0 (Applied Biosystems).

Sequences were manually aligned using the ssrQrZAG30 reference sequence (kindly provided by H. Steinkellner, Center of Applied Genetics, Vienna, Austria). In those alleles, in which we detected an unexpected relative difference of several base pairs between fragment and sequence length, we assumed that a stutter fragment was inserted in the colony and hence manually added the respective repeat copies (maximum four copies; underlined in Fig. 2).

We performed an unweighted pair group method with arithmetic mean (UPGMA) cluster analysis from a $0 / 1$ matrix in which we equally weighted nucleotide substitutions and indels but excluded the GA repeat region (software TFPGA; http:// www.marksgeneticsoftware.net/tfpga.htm; 1,000 bootstrap replicates).

## Results

Allele sequences showed three types of sequence variation across the consensus sequence length of 247 bp (excluding priming sites; Fig. 2). (1) Alleles differed by


Fig. 2 Alignment of partial allele sequences at locus ssrQrZAG30 (Kampfer et al. 1998) of oaks (Quercus spp.; GenBank accession numbers DQ323460-DQ323498) with reference sequence (Ref.; provided by H. Steinkellner, Center of Applied Genetics, Vienna). The box marks the microsatellite repeat region. Sequences are ordered according to three allele size classes and to repeat copy number. Numbers in the top line indicate positions relative to the consensus sequence of 247 bp . Nucleotide substitutions (bold) and manually added GA repeat copies based on fragment length analysis (underlined) are highlighted. Two and three different alleles originally assigned to the same allelic bin, but representing cases of homoplasy, are indicated by crosses and asterisks, respectively. $p, Q$. petraea; $r, Q$. robur; $u$, unclassified (putative hybrid; Gugerli et al. 2007); slashes indicate omitted sequence stretches; dashes indicate gaps
multiples of 2 bp owing to repeat copy number variation, (2) few nucleotide substitutions occurred within the repeat motif $(G \rightarrow A, A \rightarrow G)$ and in the flanking regions, and (3) indels were detected in the flanking region downstream of the repeat region. The two longest indels ( 33 and 28 bp ) were tandemly positioned and characterized three classes of alleles (Fig. 2). The small allele size class further differed from the medium and large classes by a nucleotide substitution (A-G) at position 129, while a nucleotide substitution (G-A) at position 240 grouped the 171-bp allele of the small allele size class with alleles of the medium and large size classes (Fig. 2). Three indels and one nucleotide substitution separated the medium class alleles from the single allele in the largest size class. Copy numbers of the repeat motif (GA) varied between seven and 32 or 35 in the small and medium allele classes, respectively, with 20 repeat copies in the largest allele (Fig. 2). An AA motif directly following the repeat region inconsistently occurred in the small allele size class (Fig. 2), which could also have been interpreted as a nucleotide substitution $(\mathrm{G} \rightarrow \mathrm{A})$ within the SSR repeat motif.

Within the small and medium allele classes, the number and intensity of the stutter fragments in FLA increased with increasing number of GA repeats in the sequence, indicative of increasing DNA polymerase slippage during PCR amplification. This gradual change led to the difference in stuttering patterns in the size range where the two allele classes overlapped in fragment lengths (201-224 bp; Fig. 1). PCR products of the small allele class showed intensive stuttering owing to high repeat copy numbers, whereas fragments of the medium class, but of similar length, mainly had one strong peak in the electropherogram because these fragments contained only few repeat copies (Fig. 1). The large allele size class contained two mononucleotide repeats, (T) ${ }_{13}$ and $(\mathrm{A})_{8}$ (Fig. 2), which resulted in a characteristic 1-bp stuttering pattern in FLA, overlaying the 2-bp stuttering typical for a dinucleotide repeat motif (Fig. 1).

We observed several cases of size homoplasy, two of which were related to compensating length effects of copy number variation and indels. Three alleles were given the same state in FLA ( 224 bp ) but belonged to either the small and medium allele classes and displayed different stuttering patterns indicative of the respective allele class (Figs. 1 and 2). The two homoplasic 224-bp alleles within the small allele class differed by two nucleotide substitutions, one in the upstream flanking region and one in the repeat motif, and by a 2-bp indel directly following the repeat region, the latter substitution compensating for the length differences owing to repeat copy number variation. Likewise, the two sequences of the 212-bp allele either had an additional GA repeat or the AA indel directly following the repeat region (Fig. 2). Two further homoplasies within allele size classes occurred in our dataset owing to nucleotide substitutions at various positions (alleles 183, 197, 257 bp; Fig. 2).

In the UPGMA tree, we identified three clusters with bootstrap values of $>50 \%$, separating the three allele classes (data not shown), but there was no indication of a phylogenetic signal with regard to the two taxa studied, $Q$. petraea and $Q$. robur.

## Discussion

Among the disadvantages of microsatellites in population genetic studies, homoplasy is often named as possibly confounding the results (Selkoe and Toonen 2006).

However, the underlying sequence variation of fragment length variation has rarely been thoroughly investigated (but see, e.g., Ortí et al. 1997; Curtu et al. 2004; Shepherd and Lambert 2005). Here, we show that stuttering pattern variation in nSSR electropherograms and its underlying sequence variation indicate size homoplasy but at the same time help in correctly assigning alleles and, thus, minimizing error rates owing to homoplasy.

In our study on nuclear microsatellite variation in two closely related and interfertile oak species, we identified large allelic variation beyond that caused by repeat copy number variation. Three classes of alleles differed in indels in the downstream flanking region, which had intertwining effects on fragment lengths. Only manual scoring of alleles, taking into account characteristic stuttering patterns of allele size classes (Fig. 1), allowed us to perceive nearly the full range of alleles present in the dataset and to separate otherwise homoplasic alleles (Fig. 2). This avoidance of size homoplasy is of particular relevance for kinship-based inferences such as pedigree or parentage analyses. We therefore stress the importance of careful allele binning, which may help to reduce biases in datasets and the respective parameter estimations (Hoffman and Amos 2005).

Variation in nSSR flanking regions is likely to be the rule rather than the exception. However, the degree of flanking sequence variation shows positional effects, with above-average mutation rates in the nucleotide stretches directly adjacent to repeat regions (5-10 bp; Brohede and Ellegren 1999; González-Martínez et al. 2004). These findings agree with our result of the AA indel (or mutation) downstream of the microsatellite region (positions 114/115; Fig. 2). One may assume that such variation is less likely around trinucleotide repeats than around other repeat types (di-, tetranucleotide, etc.), given that the former repeat type more frequently occurs in conserved coding (Ellegren 2004) or promoter regions. Sequence variation in flanking regions has also been used for phylogenetic inference (Ortí et al. 1997). Counterintuitively, these authors found that the mutation rate in flanking sequences is higher than in repeat regions, which they take as an argument against analyses relying on a stepwise mutation model (see below).

Our observations have several implications. First, the question arises of what we define as a "microsatellite" locus, i.e., how should variation in the repeat region be delimited from variation in the flanking sequences. If only the variation in repeat copy number had been considered at the locus studied, e.g., by anchoring primers in the repeat region, our dataset would have been strongly affected by size homoplasy. Furthermore, null alleles would have been highly likely owing to the indel of two bases (AA) at the $3^{\prime}$-end binding site of the reverse primer (positions 114-115; Fig. 2). Excess variation directly adjacent to the repeat region has been shown also for other organisms and may be considered characteristic for microsatellite flanking regions (Brohede and Ellegren 1999).

Second, implementing stuttering pattern variation in algorithms for automatic allele calling would be desirable but is not straightforward, given that stuttering patterns may also vary as a consequence of low primer hybridization or when the two alleles in a heterozygote individual are of similar length with overlapping stutter fragments (Selkoe and Toonen 2006). Alternatively, a combination of traditional FLA with single-strand conformational polymorphism (SSCP) analysis would be required for fully resolving the allelic diversity due to substitution-based homoplasy.
Springer

Third, variation outside the microsatellite region affects statistical analyses based on a stepwise mutation model (SMM; Kimura and Ohta 1978), though Adams et al. (2004) found only minor effects of size homoplasy on population genetic estimates using SMM.

In our study, the alleles of all three size classes were shared between the two oak taxa studied (Fig. 2; Gugerli et al. 2007). Hence, flanking region variation was phylogenetically uninformative. A similar observation was reported by Curtu et al. (2004) for another nSSR locus developed for oaks (Steinkellner et al. 1997). At the same time, sequence variation at the ssrQrZAG30 locus does not provide any information that could help to resolve the highly debated issue of shared ancestral genetic polymorphism or recurrent gene flow (hybridization) between the studied white oak taxa Q. petraea and Q. robur (Lexer et al. 2006; Muir and Schlötterer 2006).

We finally conclude that SSR loci need to be carefully evaluated during their development by sequencing a large number of alleles and by confirming their Mendelian inheritance (Reece et al. 2004; Selkoe and Toonen 2006). Selecting appropriate primer sites is crucial not only with regard to amplification success but primer positions may also have strong implications on the feasibility of automatic allele scoring and on the level of size homoplasy in a dataset. However, indel-based variation in flanking regions may leave its imprint in the FLA electropherograms, like in the stuttering pattern variation presented here, which may be helpful in correct allele calling.

Acknowledgements This work was carried out in the framework of the European project Oakflow (CTFAIR 2000-00960), which was financially supported by the Swiss Federal Office for Education and Science (BBW 99.0838). We thank Christoph Sperisen for providing technical support and an anonymous referee for helpful comments on the manuscript.

## References

Adams RI, Brown KM, Hamilton MB. The impact of microsatellite electromorph size homoplasy on multilocus population structure estimates in a tropical tree (Corythophora alta) and an anadromous fish (Morone saxatilis). Mol Ecol. 2004;13(9):2579-88.
Brohede J, Ellegren H. Microsatellite evolution: polarity of substitutions within repeats and neutrality of flanking sequences. Proc R Soc Lond B. 1999;266(1421):825-33.
Curtu A-L, Finkeldey R, Gailing O. Comparative sequencing of a microsatellite locus reveals size homoplasy within and between European oak species (Quercus spp.). Plant Mol Biol Reptr. 2004;22 (4):339-46.

Ellegren H. Microsatellites: simple sequences with complex evolution. Nat Rev Genet. 2004;5(6):435-45.
Estoup A, Jarne P, Cornuet J-M. Homoplasy and mutation model at microsatellite loci and their consequences for population genetics analysis. Mol Ecol. 2002;11(9):1591-604.
Goldstein DB, Schlötterer C. Microsatellites: evolution and application. Oxford: Oxford University Press; 1999.

González-Martínez SC, Robledo-Arnuncio JJ, Collada C, et al. Cross-amplification and sequence variation of microsatellite loci in Eurasian hard pines. Theor Appl Genet. 2004;109(1):103-11.
Gugerli F, Walser J-C, Dounavi K, et al. Coincidence of small-scale spatial discontinuities in leaf morphology and nuclear microsatellite variation of Quercus petraea and $Q$. robur in a mixed forest. Ann Bot. 2007;99(4):713-22.
Hoffman JI, Amos W. Microsatellite genotyping errors: detection approaches, common sources and consequences for paternal exclusion. Mol Ecol. 2005;14(2):599-612.

Kampfer S, Lexer C, Glössl J, et al. Characterization of (GA)(n) microsatellite loci from Quercus robur. Hereditas 1998;129(2):183-6.
Kimura M, Ohta T. Stepwise mutation model and distribution of allelic frequencies in a finite population. Proc Natl Acad Sci U S A. 1978;75(6):2868-72.
Lexer C, Kremer A, Petit RJ. Shared alleles in sympatric oaks: recurrent gene flow is a more parsimonious explanation than ancestral polymorphism. Mol Ecol. 2006;15(7):2007-12.
Muir G, Schlötterer C. Moving beyond single-locus studies to characterize hybridization between oaks (Quercus spp.). Mol Ecol. 2006;15(8):2301-4.
Ortí G, Pearse DE, Avise JC. Phylogenetic assessment of length variation at a microsatellite locus. Proc Natl Acad Sci U S A. 1997;94(20):10745-9.
Reece KS, Ribeiro WL, Gaffney PM, et al. Microsatellite marker development and analysis in the Eastern Oyster (Crassostrea virginica): confirmation of null alleles and non-Mendelian segregation ratios. J Hered. 2004;95(4):346-51.
Schlötterer C. Genome evolution: are microsatellites really simple sequences. Curr Biol. 1998;8(4):R132-4.
Selkoe KA, Toonen RJ. Microsatellites for ecologists: a practical guide to using and evaluating microsatellite markers. Ecol Lett. 2006;9(5):615-29.
Shepherd LD, Lambert DM. Mutational bias in penguin microsatellite DNA. J Hered. 2005;96(5):566-71.
Steinkellner H, Fluch S, Turetschek E, et al. Identification and characterization of (GA/CT)n-microsatellite loci from Quercus petraea. Plant Mol Biol. 1997;33(6):1093-6.
Zhang D-X, Hewitt GM. Nuclear DNA analyses in genetic studies of populations: practice, problems and prospects. Mol Ecol. 2003;12(3):563-84.


[^0]:    F. Gugerli $(\boxtimes) \cdot$ S. Brodbeck $\cdot$ R. Holderegger

    Ecological Genetics and Evolution, WSL Swiss Federal Research Institute, Zürcherstrasse 111, 8903 Birmensdorf, Switzerland
    e-mail: gugerli@wsl.ch

[^1]:    Springer

