Molecular Ecology (2004) **13**, 2101–2106 doi: 10.1111/j.1365-294X.2004.02209.x

SHORT COMMUNICATION

A simple method for the calculation of microsatellite genotype distances irrespective of ploidy level

RUŽICA BRUVO, NICOLAAS K. MICHIELS, THOMAS G. D'SOUZA and HINRICH SCHULENBURG
Department of Evolutionary Biology, Institute for Animal Evolution and Ecology, Westphalian Wilhelms-University, Hüfferstr. 1, 48149 Münster, Germany

Abstract

Microsatellites are powerful molecular markers, used commonly to estimate intraspecific genetic distances. With the exception of band sharing similarity index, available distance measures were developed specifically for diploid organisms and are unsuited for comparisons of polyploids. Here, we present a simple method for calculation of microsatellite genotype distances, which takes into account mutation processes and permits comparison of individuals with different ploidy levels. This method should provide a valuable tool for intraspecific analyses of polyploid organisms, which are widespread among plants and some animal taxa. An illustration is given using data from the planarian flatworm *Schmidtea polychroa* (Platyhelminthes).

Keywords: microsatellites, parentage, polyploids, population, relatedness, *Schmidtea*

Received 12 November 2003; revision received 26 February 2004; accepted 19 March 2004

Introduction

Microsatellites are used widely for the analysis of variation within and between populations. One application includes the calculation of genetic distances between individuals in order to infer levels of relatedness or to determine parentage, which often represent an important basis for subsequent evolutionary or ecological analysis of phenotypical traits (e.g. Blouin 2003; Jones & Ardren 2003). The distance calculations are based usually on the proportion of shared alleles (Lynch 1990; Bowcock *et al*. 1994). This method follows implicitly the infinite alleles model, thus assuming independence of alleles and ignoring mutational processes, which can result in biased distances especially when alleles are highly polymorphic.

Alternative approaches are based on the stepwise mutation model and extensions thereof, which incorporate a higher likelihood for small than for large changes in microsatellite repeat number. For instance, Otter *et al*. (2001, 2003) suggested calculation of genetic distances between individuals as the squared number of repeat differences among pairwise compared alleles. Their approach is based on the microsatellite evolution model by Goldstein *et al*. (1995), and it is related to those developed by Slatkin (1995) and Streiff *et al*. (1998). However, the squaring of repeat number differences may overemphasize distances between the alleles that are several repeat units apart. In contrast, several additional models on microsatellite evolution derive the likelihood of allele size changes from a symmetric geometric distribution as a function of the microsatellite mutation rate (Di Rienzo *et al*. 1994; Fu & Chakraborty 1998; Pritchard *et al*. 1999; Slatkin 2002). In this case, the rate of decrease in likelihood for large changes does not amplify as above, but it diminishes with increasing size differences. This latter approach has not been employed as yet for the inference of genetic distances between individuals. Furthermore, the currently available distance calculation methods were devised for diploid organisms. Hence, they are not suited for the analysis of organisms with higher ploidy levels.

In such comparisons, two common problems arise: (i) certain species may combine individuals of different ploidy classes. In these cases, the microsatellite allele distribution bears valuable information about levels of relatedness, but it cannot be evaluated as yet using existing methods. (ii) The number of detected different alleles in a polyploid individual is often lower than its ploidy level. For instance, a triploid individual may have alleles of only

Correspondence: Hinrich Schulenburg. Fax: + 49 251 8324668; E-mail: hschulen@uni-muenster.de

two different sizes, a and b. If the presence of null alleles is ruled out, then either allele a or b occurs twice, i.e. genotypes aab or abb. In the following, we refer to this state as 'partial heterozygotes'.

Here we present a new and simple method for calculating relative distances between microsatellite genotypes, which specifically permits analysis of polyploids and which takes into account stepwise mutational processes. We illustrate the principle of our method using data from a polyploid population of the planarian flatworm *Schmidtea polychroa* (Platyhelminthes), in which parthenogenetic triploid and tetraploid individuals coexist. Individuals of different ploidy levels are related to each other, because asexual polyploids repeatedly originate from sexual diploids (Weinzierl *et al*. 1999), and because different asexual polyploid lineages occasionally exchange genetic material (i.e. occasional sex; D'Souza *et al*. 2004).

Methods

The model

We assume that slipped-strand mispairing is the main cause of changes in microsatellite length, that it usually causes single-step mutations, and that changes involving two, three and more steps occur with diminishing frequency. Thus, the probability m of occurrence of a specific change in repeat number decreases exponentially with the number of repeat unit differences or steps, x, as predicted by the extended stepwise mutation model, where m is calculated from a symmetric geometric distribution (Di Rienzo *et al*. 1994; Fu & Chakraborty 1998; Pritchard *et al*. 1999; Slatkin 2002). For our calculations, we use a specific case of this model, namely:

$$m_x = 2^{-|x|} \qquad \text{(eqn 1)}$$

This specific case is a simplification of the original model in that it ignores the microsatellite mutation rate, which is usually unknown and which is not required for intrapopulational comparison of relative genotype distances inferred from identical sets of microsatellite loci. Moreover, we use the median value for the fraction of single-step changes, which are usually also unknown in empirical studies. Most importantly, employment of the median value ascertains that the model takes the form given in equation 1, which represents the only case, for which all m_x values lie in between 0 and 1, thus permitting simple calculation of the genetic distance between two alleles, d_a, as:

$$d_a = 1 - m_x \qquad \text{(eqn 2)}$$

Following the parsimony principle, the distance d_l between two single locus genotypes is then given as the minimum sum of all between-allele comparisons divided by the ploidy level, k:

$$d_l = \frac{\left(\min_{i=1\ldots k!} s_i\right)}{k} \qquad \text{(eqn 3)}$$

where s_i is the sum of a compatible set of allele comparisons for a specific locus:

$$s_i = \sum_{a=1}^{k} d_a \qquad \text{(eqn 4)}$$

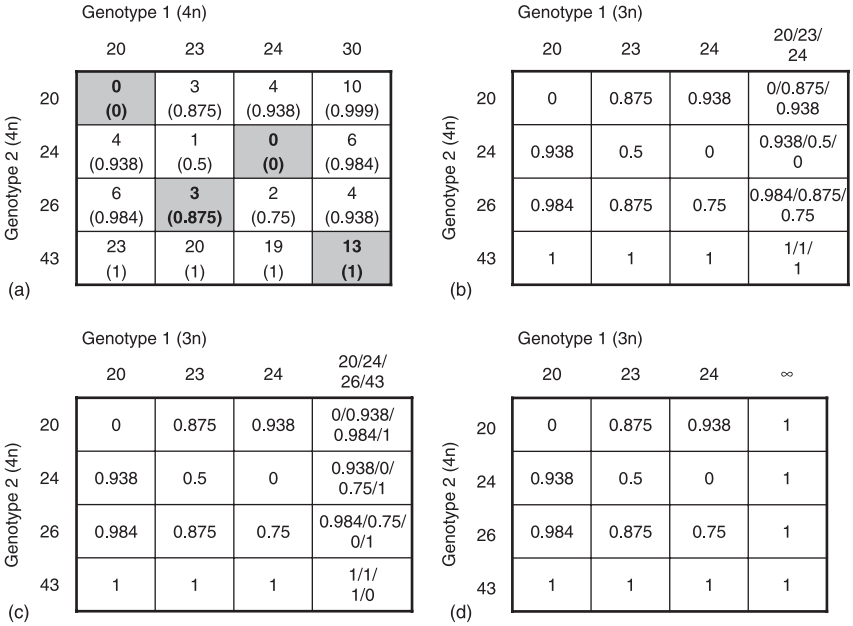
The overall distance between two individuals, D, is subsequently obtained by the sum of the locus-distances divided by the number of loci, l:

$$D = \frac{\sum_{i=1}^{l} d_l}{l} \qquad \text{(eqn 5)}$$

For illustration, consider two tetraploid individuals each with four different alleles at one microsatellite locus. Alleles are represented as the number of repeat units. In this case, calculations are performed using a 4×4 matrix, with alleles arranged on the margins and cells within the matrix representing the 'steps' between them (Fig. 1A). The genetic distances between all different allele combinations are then calculated with d_a (Fig. 1A). Thereafter, we estimate the minimum distance: first we calculate all the possible sums of four values from the matrix by taking each time only one value from a row and one value from a column (s_i in equations 3 and 4). In this way, each allele from individual 1 is compared to only one allele from individual 2 at a time. For a 4×4 matrix, there are $4! = 24$ such combinations. Then we look for the combination yielding the smallest sum. In our example, this sum is $s = 0 + 0.875 + 0 + 1 = 1.875$ (grey shading in Fig. 1A). The result is standardized by dividing by the ploidy level, such that $d_l = 1.875/4 = 0.469$. This relative distance ranges from zero for identical genotypes to 1 for the totally different ones. The overall distance is then simply calculated by averaging d_l across all loci.

Special cases

For distance estimates between genotypes with different ploidy levels we suggest the following approach. The genotype with the lower ploidy is extended to the level of the other genotype by adding 1 or more virtual allele(s) at each locus under consideration. Each virtual allele is then given one or several values. These values are chosen specifically to 'simulate' the most likely mechanism for the occurrence of ploidy differences. For instance, such



(a) Genotype 1 (4n) × Genotype 2 (4n)

Genotype 2 (4n) \ Genotype 1 (4n)	20	23	24	30
20	**0 (0)**	3 (0.875)	4 (0.938)	10 (0.999)
24	4 (0.938)	1 (0.5)	**0 (0)**	6 (0.984)
26	6 (0.984)	**3 (0.875)**	2 (0.75)	4 (0.938)
43	23 (1)	20 (1)	19 (1)	**13 (1)**

(b) Genotype 1 (3n) × Genotype 2 (4n)

Genotype 2 (4n) \ Genotype 1 (3n)	20	23	24	20/23/24
20	0	0.875	0.938	0/0.875/0.938
24	0.938	0.5	0	0.938/0.5/0
26	0.984	0.875	0.75	0.984/0.875/0.75
43	1	1	1	1/1/1

(c) Genotype 1 (3n) × Genotype 2 (4n)

Genotype 2 (4n) \ Genotype 1 (3n)	20	23	24	20/24/26/43
20	0	0.875	0.938	0/0.938/0.984/1
24	0.938	0.5	0	0.938/0/0.75/1
26	0.984	0.875	0.75	0.984/0.75/0/1
43	1	1	1	1/1/1/0

(d) Genotype 1 (3n) × Genotype 2 (4n)

Genotype 2 (4n) \ Genotype 1 (3n)	20	23	24	∞
20	0	0.875	0.938	1
24	0.938	0.5	0	1
26	0.984	0.875	0.75	1
43	1	1	1	1

Fig. 1 Illustration of the 4×4 matrix used for the calculations of microsatellite distances among polyploid individuals. (A) Calculation of the absolute number of repeat length differences (top values) and the genetic distance (bottom values in brackets) between alleles of a single microsatellite locus from two tetraploid genotypes. Genetic distances were calculated as $d_a = 1 - 2^{-|x|}$, where x represents the number of repeat differences. The cells, which produce the minimum sum of a compatible set of alleles, are highlighted in grey. (B and C) Genetic distances between pairwise compared alleles between a triploid and a tetraploid genotype, using the same formula as in (A), whereby ploidy differences are assumed to result from either genome addition (B; virtual allele = real alleles present in the same genotype) or genome loss (C; virtual allele = real alleles present in the second genotype of the comparison). (D) Simplified procedure for genetic distance calculations between a triploid and a tetraploid genotype. The virtual allele is set to infinity, followed by distance calculations as in (A). In all cases, the margins give the repeat number of the different alleles and the cells show the differences between them.

differences may be caused with equal likelihood by genome additions from the same individual (i.e. autopolyploidization) or genome loss. In this case, to take account of genome addition, the virtual allele is first given the value of any of the real alleles present in the same genotype, followed by calculation of the average d_l estimate. Thereafter, to take account of genome loss, the virtual allele is given the value of any of the real alleles in the second genotype, again followed by calculation of the average d_l estimate. The final distance is then calculated as the mean of the two alternatives, which ensures that both have the same likelihood. The result is subsequently standardized by dividing by the higher of the two ploidy levels:

$$d_{l[\text{different ploidies}]} = \frac{\bar{d}_{l[\text{genome addition}]} + \bar{d}_{l[\text{genome loss}]}}{2k} \qquad \text{(eqn 6)}$$

This procedure is illustrated in Fig. 1B, and C. Here, $\bar{d}_{l[\text{genome addition}]} = (1.875 + 1.875 + 1.75)/3 = 1.833$, and $\bar{d}_{l[\text{genome loss}]} = (1.875 + 1.75 + 1.75 + 0.875)/4 = 1.563$, such that $d_{l[\text{different ploidies}]} = (1.833 + 1.563)/(2 * 4) = 0.424$.

The above procedure is flexible, and can be adjusted easily to alternative mechanisms, which generate differences in ploidy level. In some cases, however, the situation is more complex. For instance, in our own model system, *S. polychroa*, ploidy changes seem to be caused by diverse mechanisms including genome loss, genome additions from other individuals in the population, the joint occurrence of genome loss and addition, and possibly autopolyploidization (D'Souza *et al*. 2004). To date, the exact diversity of mechanisms and their respective frequency is unknown. If ploidy changes are relatively rare in these cases, then we propose using the following simplification. Here, each virtual allele is given a value of $x = \infty$. As a result, all comparisons including the lacking alleles are equal to 1, the highest possible distance (Fig. 1D). As above, the result is then standardized by dividing by the higher of the two ploidy levels, such that $d_l = (0 + 0 + 0.875 + 1)/4 = 0.469$. In this case, zero distances between single locus genotypes are possible only when ploidy levels are identical. For example, the pairwise distance between a triploid that shares all its alleles with a tetraploid is 0.25, not 0.

For partial heterozygotes (triploid aab or abb, etc.), the above distance calculations are performed for each possible genotype pattern and then averaged. The analogous procedure can also be used for comparison of triploids vs. diploids, tetraploids vs. diploids, and so on.

Distances based on band sharing

For a comparison between methods, we also calculate distances based on band sharing (similarity index) following the original approach (Lynch 1990). In this case, the probability of all changes is set equal to 1. The shortest distance (min s_i) is then calculated as above, and again corrected by dividing by the higher of the two ploidy levels.

Microsatellite analysis of planarian flatworms

One hundred and five parthenogenetic flatworms *S. polychroa* were collected from Lake Caldonazzo, Northern Italy. DNA was isolated with the Nucleon BACC1 DNA extraction kit (Amersham™), as described in Pongratz *et al*. (2001). Three polymorphic microsatellite loci (ATT-repeats) were amplified, SpATT12, SpATT18 and SpATT20. For SpATT20, primers, PCR conditions and analysis followed the methods described in Pongratz *et al*. (2001). For SpATT12, new primers were used (forward: 5′-CGGTTAGATTTTGC-TGGATGA-3′; reverse: 5′-GGAATGGAACGGATATT-TTAGG-3′), with the same cycle as in Pongratz *et al*. (2001). For SpATT18, new primers (forward: 5′-CGCAACAAAA-TGCTTAAATTATC-3′; reverse: 5′-TATTGGTAAAATC-TCTTGAACAAAC-3′) and a new cycling profile (2 min at 95 °C followed by 35 cycles of 20 s at 95 °C, 1 min at 60 °C for annealing and 1:30 min at 72 °C, and a final elongation step of 10 min at 72 °C) were used. Fragment analysis was as in Pongratz *et al*. (2001). Pairwise distances between multilocus genotypes were calculated using the two methods described above. The obtained distance matrices then served to construct intrapopulational genotype networks with the MINIMUM SPANNING NETWORK software (Excoffier 1993).

Results and discussion

Our method provides a new approach for the calculation of microsatellite genotype distances irrespective of ploidy levels. We consider this method to produce more realistic results than previous approaches, e.g. based on band-sharing, because it does take into account microsatellite mutation events. Allowing repeat changes to diminish in frequency with the number of steps is in agreement with empirical data on microsatellite mutations, e.g. in pipefish, chickpea, humans or passerine birds (Jones *et al*. 1999; Udupa & Baum 2001; Brohede *et al*. 2002; Huang *et al*. 2002; Beck *et al*. 2003; see also review by Ellegren 2000). Generally, these data suggest that changes of more than four steps are all extremely rare, and their occurrence may not differ much in likelihood. Mechanistically, this may be explained by the occurrence of different, but similarly rare multistep mutation events, each of which produces allele size changes of different length with essentially identical small likelihood (Kruglyak *et al*. 1998; Ellegren 2000; Schlötterer 2000). This argues clearly against between-individual distance calculations based on the squaring of repeat number differences (Otter *et al*. 2001, 2003), because squaring inflates the weight given to large-step changes. Instead, the probabilities of different mutation events are described more effectively by a symmetric geometric distribution, as realized in the extended stepwise mutation model (Di Rienzo *et al*. 1994; Fu & Chakraborty 1998; Pritchard *et al*. 1999; Slatkin 2002), which forms the basis for our approach.

Our method also solves many problems associated with distance calculations of polyploids, as it allows a versatile comparison of partial heterozygotes as well as organisms with different ploidy levels. Our method is also flexible in that it allows incorporation of prior knowledge on the mechanism of ploidy changes. If such information is not available, then a simplification is proposed for those cases, where ploidy changes are comparatively rare. This latter approach was applied to one of our own model organisms, the planarian flatworm *S. polychroa*, in which diverse mechanisms appear to produce ploidy changes (D'Souza *et al*. 2004) and for which a detailed inference of intra-populational relationships was previously impossible. Here, analysis of three microsatellite loci in 105 parthenogenetic polyploid flatworms allowed identification of 57 different multilocus genotypes (41 triploid and 16 tetraploid). These could be used to reconstruct a genotype network (Fig. 2). Consideration of microsatellite mutation processes results clearly in a different network (Fig. 2B vs. 2A). Interestingly, the network accounting for microsatellite evolution shows much higher resolution and contains distinct genotype clusters (Fig. 2B), as expected for the predominantly asexually reproducing biotypes of *S. polychroa* considered here. It is worth noting that essentially identical results were obtained when — for comparative purposes — we repeated these calculations assuming that ploidy changes result only from genome addition (i.e. autopolyploidization) and genome loss, both with equal likelihood (results not shown). This suggests that our method may be influenced to only a minor extent by the exact assumptions about ploidy changes.

In conclusion, the availability of this method should be of great value for a large diversity of model systems, which include polyploids and species with ploidy variation. These are particularly common in plants (e.g. De Kovel & De Jong 2000; Van der Hulst *et al*. 2003) and some animal taxa, including snails and crustaceans (Adamowicz *et al*. 2002; Weetman *et al*. 2002; see also review by Otto & Whitton 2000). For these systems, our method should permit a more realistic inference of relatedness or parentage, and it should thus prove useful in a variety of evolutionary and ecological studies, which aim at relating individual relationships to associated phenotypic traits. In the future, it would be desirable to incorporate our approach into a population analysis

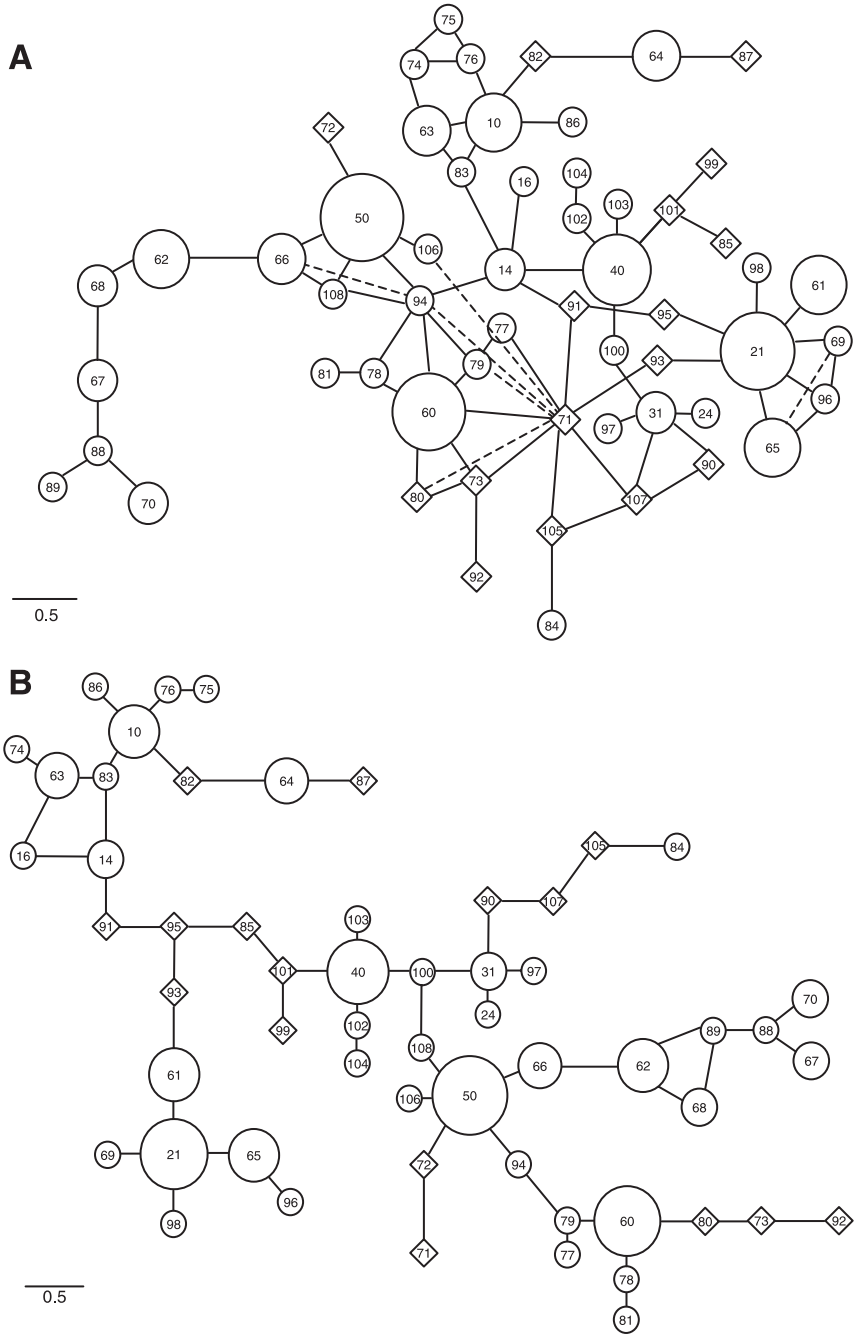


Fig. 2 Network for 57 different microsatellite genotypes of *Schmidtea polychroa* inferred from distances based on (A), band sharing, or (B), our new method. Both networks were generated using the program MINIMUM SPANNING NETWORK (Excoffier 1993). Genotypes are represented as circles (triploids) or diamonds (tetraploids) and labelled with arbitrary numbers. Circle sizes correspond to the number of individuals per genotype. Branch lengths are given in proportion to the inferred genetic distance between genotypes (see bar in bottom left corner). In (A), dotted lines are used when it was not possible to depict branch lengths correctly due to relative positions of genotypes within the network.

framework, which would allow inference of population structure, growth or migration rates, but for which the currently available programs cannot account for polyploid individuals or populations with differences in ploidy levels.

Acknowledgements

We thank Barbara Hasert for laboratory assistance and Wibke Bischoff, Gregor Schulte and Christian Wetzel for help in collecting flatworms. This work is supported by a research grant from the German Science Foundation (grant MI 482/5–1).

References

Adamowicz SJ, Gregory TR, Marinone MC, Hebert PDN (2002) New insights into the distribution of polyploid *Daphnia*: the Holarctic revisited and Argentina explored. *Molecular Ecology*, **11**, 1209–1217.

Beck NR, Double MC, Cockburn A (2003) Microsatellite evolution at two hypervariable loci revealed by extensive avian pedigrees. *Molecular Biology and Evolution*, **20**, 54–61.

Blouin MS (2003) DNA-based methods for pedigree reconstruction and kinship analysis in natural populations. *Trends in Ecology and Evolution*, **18**, 503–511.

Bowcock AM, Ruizlinares A, Tomfohrde J *et al.* (1994) High-resolution of human evolutionary trees with polymorphic microsatellites. *Nature*, **368**, 455–457.

Brohede J, Primmer CR, Moller A, Ellegren H (2002) Heterogeneity in the rate and pattern of germline mutation at individual microsatellite loci. *Nucleic Acids Research*, **30**, 1997–2003.

D'Souza TG, Storhas M, Schulenburg H, Beukeboom LW, Michiels NK (2004) Occasional sex in an 'asexual' polyploid hermaphrodite. *Proceedings of the Royal Society London Series B*, **27**, 1001–1007.

De Kovel CGF, De Jong G (2000) Selection on apomictic lineages of *Taraxacum* at establishment in a mixed sexual-apomictic population. *Journal of Evolutionary Biology*, **13**, 561–568.

Di Rienzo A, Peterson AC, Garza JC *et al.* (1994) Mutational processes of simple-sequence repeat loci in human-populations. *Proceedings of the National Academy of Sciences USA*, **91**, 3166–3170.

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

Excoffier L (1993) *Minimum Spanning Network*. Genetics and Biometry Laboratory, Department of Anthropology, University of Geneva, Geneva.

Fu Y-X, Chakraborty R (1998) Simultaneous estimation of all the parameters of a stepwise mutation model. *Genetics*, **150**, 487–497.

Goldstein DB, Linares AR, Cavallisforza LL, Feldman MW (1995) An evaluation of genetic distances for use with microsatellite loci. *Genetics*, **139**, 463–471.

Huang QY, Xu FH, Shen H *et al.* (2002) Mutation patterns at dinucleotide microsatellite loci in humans. *American Journal of Human Genetics*, **70**, 625–634.

Jones AG, Ardren WR (2003) Methods of parentage analysis in natural populations. *Molecular Ecology*, **12**, 2511–2523.

Jones AG, Rosenqvist E, Berglund A, Avise JC (1999) Clustered microsatellite mutations in the pipefish *Syngnathus typhle*. *Genetics*, **152**, 1057–1063.

Kruglyak S, Durrett RT, Schug MD, Aquadro CF (1998) Equilibrium distributions of microsatellite repeat length resulting from a balance between slippage events and point mutations. *Proceedings of the National Academy of Sciences USA*, **95**, 10774–10778.

Lynch M (1990) The similarity index and DNA fingerprinting. *Molecular Biology and Evolution*, **7**, 478–484.

Otter KA, Murray B, Holschuh C (2003) Measuring allelic variability between individuals using microsatellites. *ISBE Newsletter*, **15**, 12–15.

Otter KA, Stewart IRK, McGregor PK *et al.* (2001) Extra-pair paternity among great tits *Parus major* following manipulation of male signals. *Journal of Avian Biology*, **32**, 338–344.

Otto SP, Whitton J (2000) Polyploid incidence and evolution. *Annual Review of Genetics*, **34**, 401–437.

Pongratz N, Gerace L, Alganza AM, Beukeboom LW, Michiels NK (2001) Microsatellite development and inheritance in the planarian flatworm *Schmidtea polychroa*. *Belgian Journal of Zoology*, **131**, 71–75.

Pritchard J, Seielstad M, Perez-Lezaun A, Feldman M (1999) Population growth of human Y chromosomes: a study of Y chromosome microsatellites. *Molecular Biology and Evolution*, **16**, 1791–1798.

Schlötterer C (2000) Evolutionary dynamics of microsatellite DNA. *Chromosoma*, **109**, 365–371.

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

Slatkin M (2002) A vectorized method of importance sampling with applications to models of mutation and migration. *Theoretical Population Biology*, **62**, 339–348.

Streiff R, Labbe T, Bacilieri R, Steinkellner H, Glössl J, Kremer A (1998) Within-population genetic structure in *Quercus robur* L. and *Quercus petraea* (Matt.) Liebl. assessed with isozymes and microsatelltes. *Molecular Ecology*, **7**, 317–328.

Udupa SM, Baum M (2001) High mutation rate and mutational bias at (TAA) (n) microsatellite loci in chickpea (*Cicer arietinum* L.). *Molecular Genetics and Genomics*, **265**, 1097–1103.

Van der Hulst RGM, Mes THM, Falque M *et al.* (2003) Genetic structure of a population sample of apomictic dandelions. *Heredity*, **90**, 326–335.

Weetman D, Hauser L, Carvalho GR (2002) Reconstruction of microsatellite mutation history reveals a strong and consistent deletion bias in invasive clonal snails, *Potamopyrgus antipodarum*. *Genetics*, **162**, 813–822.

Weinzierl RP, Schmidt P, Michiels NK (1999) High fecundity and low fertility in parthenogenetic planarians. *Invertebrate Biology*, **118**, 87–94.

This study is part of a project on the importance of parasites and the accummulation of deleterious mutations in the maintenance of sex in natural populations of the planarian flatworm *Schmidtea polychroa*. The project is carried out at the Department of Evolutionary Biology in Muenster, Germany, where we also study the evolution of hermaphrodite mating systems using flatworms, earthworms, sea slugs and the evolutionary dynamics of parasite–host interactions in the nematode *C. elegans* and earthworms. Ružica Bruvo and Thomas D'Souza are graduate students, Nicolaas Michiels is a professor and leader of the group and Hinrich Schulenburg is a research assistant.
