LIMNOLOGY
and
OCEANOGRAPHY: METHODS

Limnol. Oceanogr.: Methods 6, 2008, 412–426

Isolation of microsatellites from unknown genomes using known genomes as enrichment templates

Florian Leese[1,2], Christoph Mayer[2], and Christoph Held[1]*
[1]Alfred Wegener Institute for Polar and Marine Research, Postbox 12 0161, D-27515 Bremerhaven, Germany
[2]Department of Animal Ecology, Evolution and Biodiversity, Ruhr-University of Bochum, D-44780 Bochum, Germany

Abstract

This study analyzes the performance of a method to isolate microsatellites from largely unknown target genomes. This reporter genome protocol (RGP) utilizes naturally occurring repeat motifs in genomes of distantly related organisms as hybridization probes. The RGP proved very successful in all 13 enrichment reactions from eight marine target species (comprising a dinoflagellate, a diatom, and six arthropod species), yielding on average 85.5% positive colonies. The RGP naturally screens for all repeat types occurring in the reporter genomes and is therefore less biased than standard protocols that typically test few, short motifs only. Using the genomes of *Mus musculus*, *Drosophila melanogaster*, and *Homo sapiens* as reporter genomes in this study, 133 different di- to hexanucleotide repeat types were obtained. Success of the RGP did not depend on overall microsatellite density in the reporter genome but increased with genetic distance between target and reporter genomes because hybridization to conserved nonrepeat regions is less frequent. Relative abundance of repeat types in the reporter genome had a significant effect on repeat type frequencies in the target library. Altogether, the RGP greatly simplifies the isolation of microsatellites from unknown genomes and makes microsatellite markers more attractive for a wide range of studies.

Introduction

Microsatellites are stretches of tandemly repeated DNA motifs typically 2-6 nucleotides long and frequently found in eukaryotic and prokaryotic genomes. Their high analytical power (unlinked, codominant, single locus, multiallelic markers) has led to their continued use (Schlötterer 2004) in population genetics (Bowcock et al. 1994; Paetkau et al. 1995; Zeller et al. 2006), forensic analyses (Balding 1999), assignment studies (Jones et al. 2003), and genetic mapping (Dib et al. 1996; Weissenbach et al. 1992). However, microsatellites are still often neglected as they suffer from the drawback that their initial setup is unpredictably difficult (e.g., Weetman et al. 2007). As primer sites are generally not conserved across species boundaries, microsatellite loci have to be identified de novo for most species. This often led researchers to choose less powerful but more easily obtainable markers such as RAPD, ISSR, mtDNA, AFLP, or RFLP. Despite a growing number of efficient microsatellite isolation protocols in recent years (Glenn and Schable 2005; Zane et al. 2002), there is a continuous urge for more efficient strategies to facilitate the establishment of microsatellite markers.

In model organisms whose genomes are almost completely known, the task of finding microsatellite loci is reduced to scanning sequence databases with a suitable search tool (Benson 1999, Phobos, http://www.rub.de/spezzoo/cm/). In more exploratory studies, focusing on organisms with almost unknown genomes, the aim is to find informative markers with as little effort as possible. There are two main strategies to achieve this (see Zane et al. 2002 for further details):

The first, more traditional strategy involves building a nonenriched, partial genomic library of the genome of the organism under study (from now on referred to as *target genome*). Those fragments in the library that include

*Corresponding author: E-mail: Florian.Leese@rub.de
phone: + 49-234-32-25072, fax: + 49-234-32-14114

Acknowledgments

The authors would like to thank Arne Nolte (Université Laval, Canada) for helpful advice and technical support, Tilman Alpermann (AWI Bremerhaven) for creating the genomic library of *A. tamutum* and *P. multistriata* and helpful technical advice, and Alec Lüdeking (Napels) for kindly providing samples of the diatom *P. multistriata.* The authors also want to thank Shane Windsor (University of Auckland, New Zealand) for linguistic improvements of an earlier version of this manuscript as well as two anonymous reviewers for helpful comments. Finally, thanks are due to Wolfgang Wägele (Museum Koenig, Bonn) for continuous support. This work was supported by the Deutsche Forschungsgemeinschaft grant HE 3391/3 to Christoph Held (AWI Bremerhaven) and in part by NSF grant OPP-0132032 to H.W. Detrich (Northeastern University, Boston, USA).

microsatellites are subsequently identified by hybridization to labeled synthetic oligomer probes consisting of typical nucleotide repeat motifs such as $(AC)_8$, $(AG)_{8,}$ etc. The major drawback of screening a nonenriched library for microsatellites is its inefficiency, i.e., many fragments without repeats must be screened to identify fragments with microsatellites.

The second, now de facto standard strategy aims to increase the efficiency by including an enrichment step, selectively favoring microsatellite-containing fragments before construction of the partial genomic library of the target genome. This is particularly important for genomes for which low densities of microsatellites have been reported. The enrichment is typically achieved by letting the repeat-containing genomic fragments hybridize to synthetic repeat oligomers and removing those without repeat elements during subsequent washing steps (Armour et al. 1994; Kandpal et al. 1994; Karagyozov et al. 1993; Kijas et al. 1994; Zane et al. 2002 for a review). This approach of enriching microsatellites has one major drawback: the microsatellite motifs searched for have to be defined a priori, introducing what we call a *selection bias* in the genomic library created. There are 4, 10, 33, 102, and 350 different and independent patterns for di-, tri-, tetra-, penta-, and hexanucleotide repeats, respectively (see *Material and procedures* for details). In an organism whose genome is unknown, the choice of synthetic repeat probes is largely arbitrary and is often based on known frequency patterns in distantly related model species. This selection bias increases the risk that even abundant repeat types in the anonymous target genome may be overlooked completely due to the choice of hybridization oligomers that are rare or absent in the target genome.

The need for highly successful protocols becomes even more obvious when taking into account that after the initial identification, often a significant proportion of the microsatellites will have to be excluded for several reasons (Fig. 1). Candidate microsatellites with excessive stutter bands, no suitable flanking region for primer placement, and either no amplicons or too many amplicons have to be discarded. The same holds true for microsatellites that would bias the subsequent data analysis steps due the presence of null alleles, allelic dropout, linkage disequilibrium, etc. (Selkoe and Toonen 2006).

In contrast to all other protocols, a novel enrichment protocol proposed by Nolte et al. (2005) does not rely on the use of synthetic probes, but uses genomes of unrelated organisms (from now on referred to as *reporter genomes*) as hybridization templates (Fig. 2). The concept behind this protocol is based on the assumption that evolutionary distant genomes will show a high similarity by state only in genomic regions of low complexity such as microsatellite loci (Levinson et al. 1985). These regions act as "non-synthetic universal hybridization probes" (Fig. 2). Consequently, the technique will preferentially detect repeat types that are abundant in both the target and the reporter genome and is thus potentially less susceptible to the effects of selection bias. Although it holds several advantages over standard protocols, its utility has not been tested systematically across different taxonomic groups since its inception (Nolte et al. 2005).

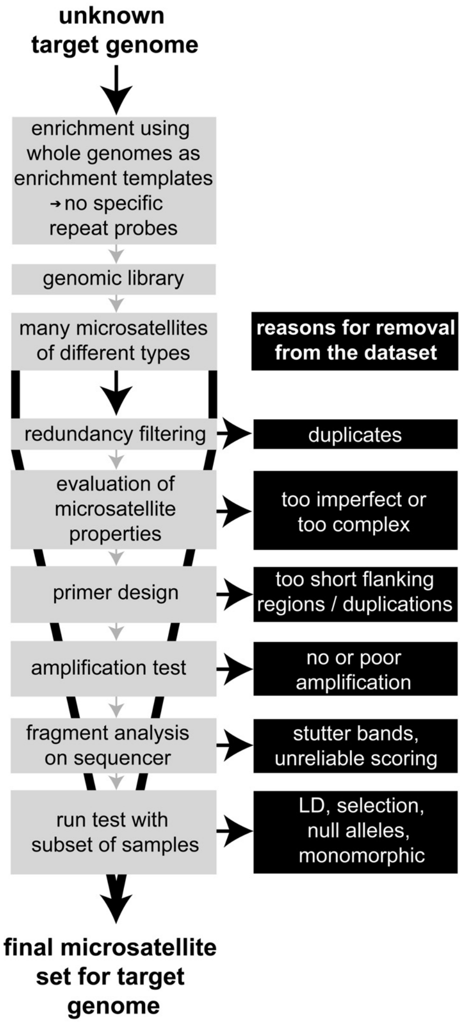
unknown target genome
enrichment using whole genomes as enrichment templates → no specific repeat probes
genomic library
many microsatellites of different types
reasons for removal from the dataset
redundancy filtering
duplicates
evaluation of microsatellite properties
too imperfect or too complex
primer design
too short flanking regions / duplications
amplification test
no or poor amplification
fragment analysis on sequencer
stutter bands, unreliable scoring
run test with subset of samples
LD, selection, null alleles, monomorphic
final microsatellite set for target genome

Fig. 1. Workflow for isolating microsatellites and steps necessary to evaluate their utility (white boxes). If rigorous standards for the quality of microsatellite markers are applied, the number of candidate loci rejected for different reasons (black boxes) often exceeds the number of loci that are eventually kept for the analysis. See Leese et al. (2008) and Kraemer et al. (in prep.) for details.

In this study, eight marine target species from three very different taxonomic groups (arthropods, diatom, dinoflagellate; Table 1) are used to systematically test the applicability and performance of the reporter genome protocol.

Material and procedures

Samples—Samples of the eight marine target species were obtained from different sources (Table 1). DNA was extracted using either the Qiagen DNeasy Tissue Mini or the Plant Mini Kit. *Mus musculus domesticus, Drosophila melanogaster* (CantonS), and *Homo sapiens* were used as reporter genomes. The *Ceratoserolis* species analyzed in this study is one species from a known complex of cryptic species, which is referred to as

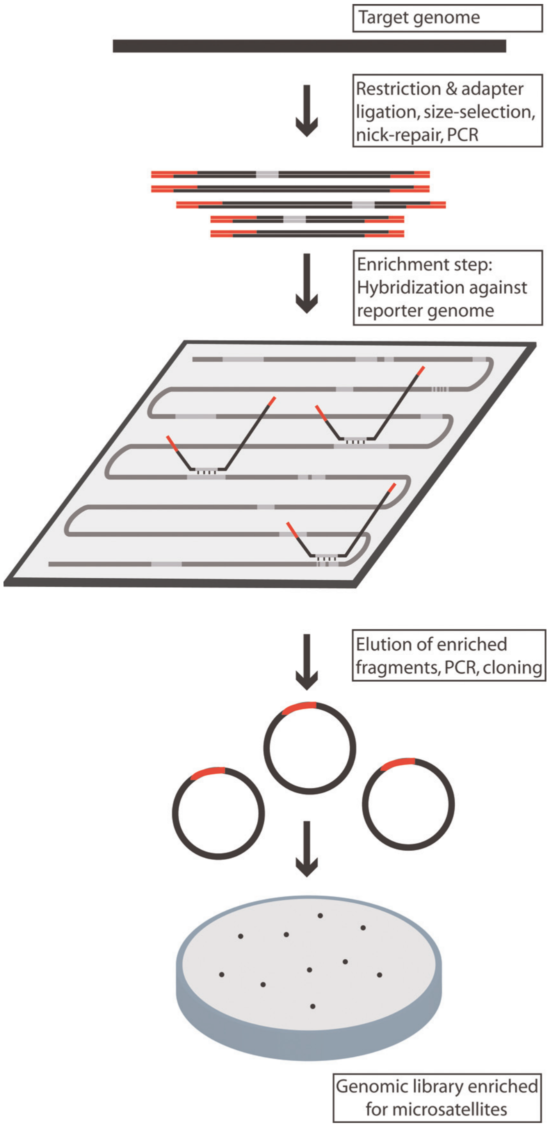
Target genome
Restriction & adapter ligation, size-selection, nick-repair, PCR
Enrichment step: Hybridization against reporter genome
Elution of enriched fragments, PCR, cloning
Genomic library enriched for microsatellites

Fig. 2. The principle of the reporter genome protocol after Nolte et al. (2005). Fragments of DNA of a target species (*target genome*) are ligated to AFLP-adaptors and hybridized against single-stranded DNA of a distantly related taxon (*reporter genome*). As a result, only fragments with a high identity-by-state are retained, predominantely simple sequence repeats (represented by light-gray areas in the DNA fragments). These fragments are eluted, cloned, and transformed into *E. coli*.

"group 1" in Held (2003) and *Ceratoserolis* n. sp. 1 in Leese and Held (2008). The species is not identical to the type species *Ceratoserolis trilobitoides sensu stricto*. A formal description is in preparation, but until its publication the species' name *C. trilobitoides* (Eights, 1833) is valid and will be used here.

Reporter genome protocol—Hybridization chips with single-stranded DNA of *Mus musculus domesticus, Drosophila melanogaster* (CantonS), and *Homo sapiens* were prepared as follows: 5 × 5 mm pieces of Hybond N+ (GE Healthcare) were incubated in a 1:1 mixture of 100 ng/µL reporter genome DNA and 1 M NaOH for 5 min at room temperature. Then the chip was transferred to 1 M Tris-CL (pH 5.0) for 2 min, following a 2-min incubation step in 2 × SSC at 50°C. Excess of liquid was removed and the DNA was baked to the membrane for 12 h at 80°C.

A small-insert genomic library enriched for microsatellites was created using total genomic DNA of the target species. Digestion of 500 ng template DNA and ligation to standard AFLP adaptors 5′-TACTCAGGACTCAT-3′/5′GACGATGAGTC-CTGAG-3′ (Vos et al. 1995) were carried out simultaneously with 20 U of the *Mse*I isoschizomer *Tru*I and 20 U of T4-DNA ligase in a 1 × Buffer-R reaction mix (50 µL; all products Fermentas Lifesciences). Both enzymes have different optimal working temperatures (65°C and 22°C). In accordance with the suggestions of the supplier, the reaction was incubated at an intermediate temperature (37°C) for 6 h, and both enzymes were added in excess. Ligase was inactivated by a 10-min hold step at 65°C. The reaction was run twice on a 1.5% TAE agarose gel, and the 400-800 bp fraction was excised and purified (Eppendorf Perfectprep® Gel cleanup). Of the 30 µL eluted DNA, 5 µL (0.5-25 ng) were amplified in a total volume of 50 µL with 0.5 µM of the AFLP adaptor-specific primer *Mse*I-N (5′-GATGAGTCCTGAGTAAN-3′) and 0.03 U/µL HotMaster® *Taq* (Eppendorf). PCR conditions were 65°C for 10 min to repair nicks remaining from adapter ligation, an initial denaturation step at 94°C for 2 min followed by 25 cycles (*P. multistriata*, *A. tamutum* 40 cycles) of 30 s at 94°C, 45 s at 52°C, 80 s at 65°C, and a final elongation step of 10 min at 65°C. The reaction was purified using the Qiaquick PCR purification Kit (Qiagen).

The hybridization chips were pre-incubated for 15 min at 50°C in 500 µL hybridization buffer (5 × SSC, 0.5% SDS, 1 × Denhardt's, 50 µg/mL Heparin). Then 400 ng template DNA was added and the tube heated to 95°C for 5 min to denature

Table 1. List of species of which microsatellite-enriched partial genomic libraries were created in this study

Species	Taxonomic position	Origin of samples
Ceratoserolis trilobitoides (Eights, 1833)	Animal, isopod crustacean	Southern Ocean, Elephant Island
Serolis paradoxa (Fabricius, 1775)	Animal, isopod crustacean	Strait of Magellan, Patagonia, Chile
Glyptonotus antarctius Eights, 1853	Animal, isopod crustacean	Southern Ocean, South Sandwich Islands
Septemserolis septemcarinata (Miers, 1875)	Animal, isopod crustacean	Southern Ocean, Bouvet Island
Munida gregaria (Fabricius, 1793)	Animal, decapod crustacean	South Atlantic Ocean, Falkland Islands
Pseudo-nitzschia multistriata Takano, 1993	Diatom algae	Provided by A. Lüdeking, Napels
Alexandrium tamutum Montresor, Beran & John, 2004	Protist, dinoflagellate	Provided by T. Alpermann, Bremerhaven

template DNA. For hybridization, the mix was kept for 30 min at 50°C. Subsequently, the hybridization chip was washed three times in hybridization buffer (50°C) and then again thrice in washing buffer (0.1 × SSC, 0.1% SDS, 50°C). Finally, DNA was eluted from the hybridization chip by transferring it into 500 μL TE buffer (pH 8.0, 5 min at 80°C). DNA was precipitated using a standard isopropanol sodium-acetate protocol (Sambrook et al. 1989).

The enriched fragments were amplified as above (25 μL reaction). Purified fragments were cloned into pCR2.1-TOPO® TA vector and transformed into competent TOP10F' or JM109 *E. coli* (Invitrogen, Promega). Cultures of positive colonies, identified by blue-white selection, were grown overnight in LB-medium containing 100 μg/mL ampicillin. Plasmid preparation was performed using the Eppendorf FastPlasmid® Mini Kit or outsourced to GATC-Biotech (Konstanz, Germany). Shotgun sequencing using standard M13-forward and/or reverse primers was either conducted in-house on an ABI 3130xl or LI-COR 4200 automated sequencer, or outsourced to Macrogen (Seoul, Korea) and GATC-Biotech (Konstanz, Germany). The addition of DMSO to a final concentration of 5% in the cycle-sequencing reaction improved the quality of reads of sequences containing microsatellites. BLAST searches were carried out for all sequences to exclude a possible contamination with mobilized DNA from the reporter genome.

In this study, eight target genomes were hybridized to some or all three reporter genomes, yielding a total of 13 libraries (see Table 2).

Nonenriched genomic library—A nonenriched partial genomic library was created for *C. trilobitoides* according to Rassmann et al. (1991). In total, 3456 clones were dotted onto four nitrocellulose membranes and screened with the radioactively labeled γ-^{33}P oligomer probes $(AC)_8$. Prehybridization and hybridization were carried out in rotating glass tubes in an incubator at 43°C. The filters and intensifying screen were exposed overnight to X-Ray films (Kodak) at –20°C and developed according to the recommendations of the manufacturer. Colonies with strong signal intensity on the film were traced back to the 96-well plates based on their position in the spatial array on the nitrocellulose membranes. Positive colonies were grown overnight, extracted (Chelex), and sequenced (see above for conditions).

Data analysis—Sequence data analysis was performed in a semi-automated workflow based on the Staden software package, version 1.70 (Staden 1996), into which the novel microsatellite search tool Phobos (Mayer, www.rub.de/spezzoo/cm) and Primer3 (Rozen and Skaletsky 2000) had been integrated (Kraemer et al. in prep.). Self-similarity dotplots were used to visualize intrinsic structural features of repeat regions using the Staden tool Spin or dotlet (Junier and Pagni 2000) and to help avoid placing primers in regions that were not unique (see Fig. 3 for examples).

First, sequence data were processed using the Staden pre-processing tool Pregap with integrated programs Phred (Ewing and Green 1998; Ewing et al. 1998) for base calling, Cross-match for cloning vector and adaptor identification, and Phobos for masking of repeats. Normal shotgun assembly of the processed reads was conducted with the Staden assembly tool Gap4, allowing for a maximum of 2% mismatch between two reads. All nonredundant contigs were entered into the Staden database for subsequent analysis. Phobos was used to search for all perfect and imperfect microsatellites present in the genomic libraries and three different reporter genomes, i.e., *Mus musculus* (NCBI Build 36.1, May 2006: ftp://ftp.ncbi.nih.gov/genomes/M_musculus, taxonomic ID 10090), *D. melanogaster* as annotated by the FlyBase consortium (FlyBase Release 5.1, February 2005; ftp://ftp.ncbi.nih.gov/genomes/Drosophila_melanogaster), and *Homo sapiens* (NCBI Build 36.2, March 2006).

Table 2. List of partial genomic libraries enriched for microsatellites using the reporter genome protocol*

Library	Target genome	Reporter genome	Unique clones	Clones with microsatellites	Number of msats	Number of different repeat types
1	*C. trilobitoides*	*M. musculus*	180	171 (95%)	395	35
2	*C. trilobitoides*	*D. melanogaster*	111	91 (82%)	226	36
3	*C. trilobitoides*	*H. sapiens*	64	62 (97%)	146	16
4	*S. paradoxa*	*M. musculus*	146	132 (90%)	351	41
5	*S. paradoxa*	*D. melanogaster*	131	110 (84%)	229	36
6	*G. antarcticus*	*M. musculus*	63	62 (98%)	179	25
7	*G. antarcticus*	*D. melanogaster*	72	61 (84%)	130	25
8	Bopyridae	*M. musculus*	28	25 (89%)	92	11
9	Bopyridae	*D. melanogaster*	45	40 (89%)	114	19
10	*S. septemcarinata*	*M. musculus*	146	37 (25%)	63	20
11	*M. gregaria*	*D. melanogaster*	9	8 (89%)	21	6
12	*Pseudo-nitzschia multistriata*	*M. musculus*	16	16 (100%)	44	7
13	*A. tamutum*	*M. musculus*	196	171 (87%)	546	88

*Given are the number of unique clones, cloning efficiency, total number, and number of different microsatellite repeat types.

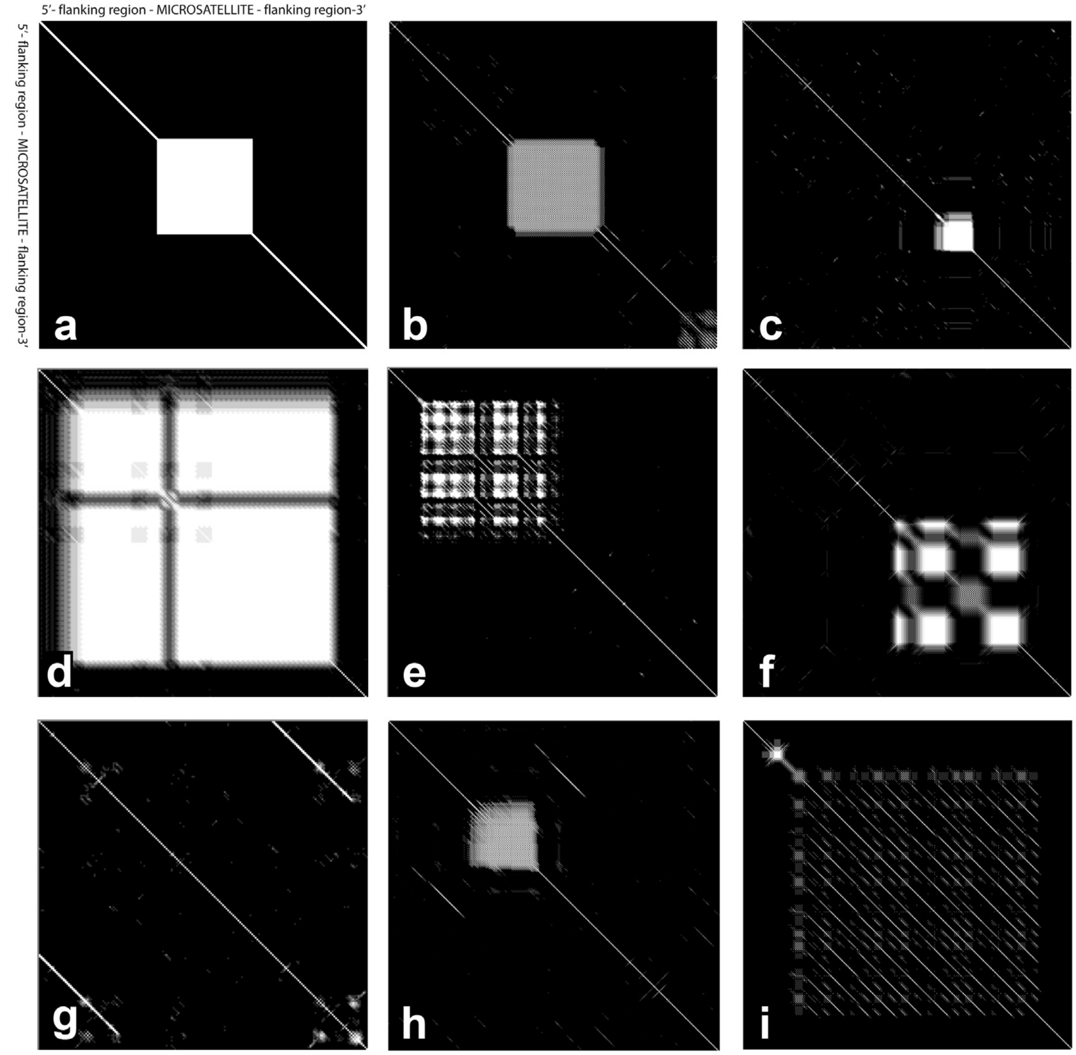

Fig. 3. Dot-plots can be used to show the self-similarity of a sequence by plotting it against itself. Bright areas indicate a high self-similarity resulting from the presence of a microsatellite (a). For population genetic analyses, highly perfect repeat loci should be preferred (b, c). Many microsatellite-containing fragments are, in fact, imperfect with one (d) or several (e) interruptions. A complex microsatellite consisting of sub-satellites with two different motifs (f). Flanking regions can contain duplications that may complicate marker setup (g, h). Loci with higher-order repeat structures (i). GenBank accession numbers: b) EU056273, c) EU056276, d) EU234060, e) EU234059, f) EU234055, g) EU234058, h) EU234056, i) EU234057.

The parameter settings used in Phobos in the present analysis were as follows: the scoring parameters (match, mismatch, N, gap) used were (1, –6, 0, –6), respectively. In each microsatellite, the first repeat unit was not scored. No more than two consecutive N's were allowed. For computing percentage perfections, N's were counted as mismatch positions. The "*recursion depth*" parameter used was 5. Microsatellites were reported if they achieved a minimum score of 8 and additionally had a minimum length of three perfect repeat units (minlength_b = 2). Consequently, only dinucleotide repeats with a minimum length of 10 bp, trinucleotide repeats with 11 bp, tetra-, penta-, and hexanucleotide repeats with a minimum length of 3 repeat units were included in the analysis. Microsatellite repeat types were reported in a canonical form according to Chambers and MacAvoy (2000), where a minimal alphabetic representation is determined by allowing cyclic permutations and computing the reverse complement of the nucleotides in the repeat motif. For instance, GA, AG, TC, CT are all represented by the repeat type AG. This convention allows us to count and identify repeat units without reference to the repeat unit phase or strand. Using this canonical representation of repeat types, as many as 4 different dinucleotide, 10 tri, 33 tetra, 102 penta, and 350 hexanucleotide repeat types exist.

After the microsatellites had been detected by Phobos, they were analyzed with the newly developed program Sat-Stat, version 1.0.1. This analysis tool is capable of filtering and sorting microsatellites according to their properties such as repeat type, unit length, total length, score, and perfection, and to analyze these properties statistically. For this study, only repeats with a unit size of 2-6 bp and percentage perfection ≥90% were considered.

To statistically assess whether a particular relative repeat class or repeat type density differs significantly among two partial genomic libraries, permutation tests were performed as follows: 100,000 randomly permuted data sets were created, each by pooling the inserts from the two original libraries and randomly redistributing them to two permuted libraries, each of which contained the same number of inserts as the original ones. Comparing the relative repeat density values in the two original libraries with the distribution of these density values in the permuted libraries allowed us to estimate the probability that the observed repeat densities could have arisen by chance from a pooled data set with no difference in the two relative densities. This test was carried out using a self-written computer program (C. Mayer, available on request).

Assessment

Reporter genome protocol—With the reporter genome protocol, 1370 clones from 13 genomic libraries and eight different target species (Table 1) were sequenced. Redundant samples were found in all but the genomic library derived from the bopyrid isopod enriched using *M. musculus* as reporter genome (∅ = 13.1%, standard deviation 13.3). All redundant sequences were removed from the analysis yielding a total number of 1207 unique sequences (Table 2).

BLASTn searches were performed to check whether sequences in the enriched library originated from the reporter genome rather than the target genome. No contaminant fragments of the reporter genome DNA were detected, indicating that mobilization of reporter genome fragments was not a problem. In the genomic library resulting from *C. trilobitoides* (target) hybridized to *D. melanogaster* (reporter), the microsatellite enrichment efficiency was reduced due to the hybridization of conserved fragments of ribosomal genes rather than fragments containing repeats. A BLAST search revealed 19 inserts originating from regions of the 18S or 28S rRNA gene, only two of these inserts also contained a microsatellite.

Efficiency of enrichment—Taking the percentage of clones with at least one microsatellite as a measure of the enrichment success, twelve of 13 genomic libraries were highly enriched for microsatellites with an average success rate of 85.5% and a standard deviation of 18.9% (Table 2). Enrichment efficiency was lowest for *S. septemcarinata* (25%) using *M. musculus* as reporter genome (library 10) and highest for *P. multistriata* and *G. antarcticus* using *M. musculus* as reporter genome (100% and 98.4%, libraries 12 and 6, Table 2). Excluding the outlier library 10, the average success rate was 90.5% ± 5.8%.

For target genomes that were enriched with more than one reporter genome, the influence of different reporter genomes on enrichment success was evaluated (libraries 1-9, Table 2). Success of the enrichment depended significantly on the reporter genome. The mean enrichment efficiency using *M. musculus* was significantly (*t*-test, $P = 0.016$) higher than using *D. melanogaster* as reporter genome (93.3% ± 4.2 and 84.9% ± 2.9, respectively). For *C. trilobitoides* as target genome, the enrichment using *H. sapiens* or *M. musculus* as reporter genomes was almost equally efficient (Table 2).

For all 13 enriched genomic libraries, 133 of 499 theoretically possible different repeat types were detected (Table 3). The major proportion of microsatellites were dinucleotide repeats, except for *P. multistriata* (Fig. 4). The proportions of longer repeat classes differed considerably between taxonomic groups (Table 3). The amount of perfect repeats ranged from 45.9% to 77.3%.

Screening 3456 colonies of the nonenriched genomic library of *C. trilobitoides* using radioactively labeled γ-^{33}P oligomer probes $(AC)_8$ yielded 26 positive colonies by colony hybridization corresponding to a success rate of 0.75%. The density of $(AC)_n$ repeats in the nonenriched library was 343 bp/Mbp and their average length was 27.5 bp. Due to the small sample size, however, these numbers represent only approximate estimates for the genomic densities and mean lengths in the genome of *C. trilobitoides*.

Influence of the characteristics of the reporter genomes on the microsatellites detected—The genomes of the three taxa used as reporter genomes differ considerably in total genome size as well as in the density, relative abundance, and lengths of their microsatellites (Table 4, Fig. 4). Density of microsatellites is highest in the genome of *M. musculus* (23733 bp/Mbp), intermediate in *D. melanogaster* (12924 bp/Mbp), and lowest in *H. sapiens* (9927 bp/Mbp).

The relative amounts of di- to hexanucleotide repeat classes and the length characteristics differ between the three genomes (Table 4, Fig. 4). Dinucleotide repeats are abundant in all three genomes, but are relatively more abundant in *M. musculus* and *H. sapiens* than in *D. melanogaster*. *D. melanogaster* is comparatively rich in trinucleotide repeats, whereas tetranucleotide repeats occur at higher densities in the genomes of *M. musculus* and *H. sapiens*. Penta- and hexanucleotide repeats constitute only a minor proportion of repeats in the genomes (Table 4).

The enriched genomic libraries (except 8, 11, and 12) contained microsatellites from all five different repeat classes (Fig. 4). However, dinucleotide repeats were the most abundant repeat class in 10 of 13 libraries.

Using a statistical permutation test (see *Material and procedures* section), it has been analyzed whether the differences in repeat type densities among the three reporter genomes used were reflected in the partial enriched genomic libraries. Specifically, we investigated whether the elevated relative density of dinucleotide repeats in the reporter genome of *M. musculus* as compared with *D. melanogaster* and the high relative density of trinucleotide repeats in *D. melanogaster* as compared with *M. musculus* (Fig. 4) lead to significant differences among

Table 3. Information on microsatellites detected in the 13 enriched libraries created using the reporter genome protocol. For each occurring repeat type the number of microsatellites and the number of inserts with at least one microsatellite of that repeat type (in brackets) are shown.

Target-Reporter Genomic library*	**Ctr-M 1**	**Ctr-D 2**	**Ctr-H 3**	**Spa-M 4**	**Spa-D 5**	**Gly-M 6**	**Gly-D 7**	**Bop-M 8**	**Bop-D 9**	**Sse-M 10**	**Mgr-D 11**	**Pmu-M 12**	**Ata-M 13**
Nr of microsatellites	395	226	146	351	229	179	130	92	114	63	21	44	546
Nr of positive inserts	171	91	62	132	110	62	61	25	40	37	8	16	171
Nr of different repeat types	35	36	16	41	36	25	25	11	19	20	6	7	88
Density [bp/Mbp]	284610	202313	265765	302243	348958	410562	347445	206999	300053	30274	231505	288330	246011
AC	69 (44)	79 (54)	48 (30)	123 (63)	91 (53)	33 (17)	33 (25)	10 (6)	34 (21)	6 (5)	6 (5)	3 (3)	179 (83)
AG	190 (110)	37 (26)	53 (31)	59 (28)	7 (4)	53 (24)	12 (7)	70 (20)	45 (12)	11 (9)	5 (3)	—	28 (22)
AT	27 (24)	11 (10)	9 (8)	23 (19)	13 (13)	16 (13)	14 (12)	2 (2)	3 (3)	11 (9)	—	—	2 (2)
CG	1 (1)	—	1 (1)	1 (1)	1 (1)	—	—	—	—	—	—	—	23 (20)
AAC	1 (1)	1 (1)	1 (1)	6 (6)	23 (23)	5 (4)	16 (11)	—	—	2 (2)	—	11 (11)	21 (12)
AAG	1 (1)	2 (1)	—	4 (3)	—	6 (5)	9 (6)	—	2 (1)	2 (2)	—	3 (3)	42 (18)
AAT	8 (6)	3 (3)	—	6 (5)	1 (1)	—	1 (1)	2 (2)	—	5 (4)	—	—	5 (4)
ACC	1 (1)	1 (1)	—	—	1 (1)	—	1 (1)	1 (1)	—	—	—	3 (3)	10 (6)
ACG	—	1 (1)	—	8 (4)	2 (2)	—	—	—	—	—	—	—	2 (2)
ACT	—	—	—	4 (4)	—	—	—	—	1 (1)	3 (2)	—	20 (9)	3 (1)
AGC	1 (1)	2 (2)	—	1 (1)	2 (2)	—	3 (1)	—	—	1 (1)	—	3 (3)	5 (4)
AGG	—	2 (1)	—	1 (1)	—	1 (1)	—	—	—	2 (2)	—	1 (1)	3 (3)
ATC	—	—	—	—	—	3 (3)	—	—	—	1 (1)	—	—	5 (4)
CCG	—	—	1 (1)	—	—	—	—	—	—	1 (1)	—	—	1 (1)
AAAC	5 (5)	3 (3)	—	4 (3)	3 (3)	1 (1)	1 (1)	—	1 (1)	—	1 (1)	—	4 (3)
AAAG	6 (6)	1 (1)	5 (5)	12 (12)	3 (3)	9 (6)	—	—	1 (1)	2 (2)	—	—	—
AAAT	2 (2)	1 (1)	—	4 (3)	—	7 (4)	—	—	—	2 (2)	1 (1)	—	—
AACC	—	1 (1)	—	—	1 (1)	—	—	—	—	—	—	—	2 (2)
AACG	—	—	—	1 (1)	—	—	—	—	—	—	—	—	1 (1)
AAGC	—	—	—	3 (3)	—	—	—	—	—	—	—	—	2 (2)
AAGG	9 (7)	1 (1)	—	7 (3)	2 (2)	—	—	1 (1)	—	—	—	—	1 (1)
AAGT	6 (6)	1 (1)	1 (1)	1 (1)	—	—	—	—	—	—	—	—	1 (1)
AATC	6 (6)	—	—	—	1 (1)	—	—	—	—	—	—	—	2 (1)
AATG	—	—	—	2 (1)	—	—	—	—	—	1 (1)	—	—	—
ACAG	29 (20)	13 (11)	14 (12)	26 (14)	23 (13)	2 (2)	—	—	5 (1)	8 (2)	—	—	12 (5)
ACAT	8 (6)	2 (2)	1 (1)	10 (8)	3 (2)	13 (11)	10 (9)	2 (2)	5 (5)	—	3 (2)	—	4 (2)
ACCG	—	—	—	—	—	—	—	1 (1)	—	—	—	—	3 (1)
ACCT	—	—	—	—	—	—	1 (1)	—	—	—	—	—	2 (2)
ACGC	—	3 (3)	—	4 (3)	2 (1)	3 (2)	1 (1)	1 (1)	2 (2)	—	—	—	39 (29)
ACGG	—	1 (1)	—	—	—	—	—	—	—	—	—	—	1 (1)
ACTC	1 (1)	1 (1)	—	3 (2)	6 (6)	1 (1)	—	—	2 (2)	—	—	—	4 (3)
ACTG	1 (1)	—	—	—	—	—	—	—	—	—	—	—	—
AGAT	3 (3)	2 (2)	2 (2)	7 (5)	5 (5)	2 (2)	2 (2)	1 (1)	4 (2)	—	—	—	—
AGCC	—	—	—	—	1 (1)	—	—	—	—	—	—	—	—
AGCG	2 (2)	—	—	—	—	—	—	—	—	—	—	—	3 (2)
AGCT	—	—	—	1 (1)	—	—	—	—	—	—	—	—	—
AGGC	—	—	—	1 (1)	—	—	—	—	—	—	—	—	1 (1)
AGGG	—	—	—	—	—	—	—	—	—	1 (1)	—	—	2 (2)
ATCC	—	—	—	1 (1)	—	—	—	—	—	—	—	—	4 (2)
ATGC	—	1 (1)	—	—	—	—	—	—	—	—	—	—	1 (1)
CCCG	—	—	—	—	—	—	—	—	—	—	—	—	1 (1)

continued

Table 3. *Continued*

AAAAC	—	—	1 (1)	—	—	—	—	—	—	—	—	—	—
AAACC	—	—	—	—	1 (1)	—	—	—	—	—	—	—	—
AAAGC	—	—	—	—	—	—	1 (1)	—	—	—	—	—	—
AACAC	—	1 (1)	—	—	—	1 (1)	—	—	—	—	—	—	4 (2)
AACCT	3 (3)	4 (4)	5 (3)	—	2 (2)	1 (1)	—	—	—	1 (1)	—	—	—
AACGC	—	—	—	—	—	—	—	—	—	—	—	—	3 (3)
AAGAC	—	15 (8)	—	—	—	—	—	—	—	—	—	—	—
AAGAG	1 (1)	22 (7)	—	—	—	1 (1)	—	—	3 (3)	—	5 (1)	—	12 (12)
AAGAT	—	—	—	—	1 (1)	—	—	—	—	—	—	—	—
AAGGC	—	—	—	—	—	—	—	—	—	—	—	—	1 (1)
AAGGG	1 (1)	—	—	—	—	—	—	—	1 (1)	—	—	—	—
AAGGT	—	—	—	—	—	—	—	—	—	—	—	—	1 (1)
AAGTC	1 (1)	—	—	—	—	—	—	—	—	—	—	—	—
AATAC	—	1 (1)	—	—	1 (1)	—	—	—	—	—	—	—	—
AATAT	—	—	—	—	—	1 (1)	—	—	—	1 (1)	—	—	—
AATCT	—	—	—	—	1 (1)	—	—	1 (1)	1 (1)	—	—	—	—
AATAG	—	2 (1)	—	—	—	—	—	—	—	—	—	—	—
AATGC	—	—	—	—	—	—	—	—	—	—	—	—	1 (1)
ACACG	—	—	—	—	—	—	1 (1)	—	—	—	—	—	—
ACAGC	—	—	—	—	—	—	—	—	—	—	—	—	10 (7)
ACATC	—	—	—	—	—	—	—	—	—	—	—	—	3 (3)
ACATG	—	—	—	—	—	—	—	—	—	—	—	—	2 (2)
ACCTC	—	—	—	—	—	—	—	—	—	—	—	—	1 (1)
ACCTG	—	—	—	—	—	—	—	—	—	—	—	—	1 (1)
ACGAG	—	1 (1)	—	—	—	—	—	—	—	—	—	—	—
ACGGC	—	—	—	—	—	—	—	—	—	—	—	—	1 (1)
ACTAG	—	1 (1)	—	—	—	—	—	—	—	—	—	—	—
AGAGC	—	—	—	—	1 (1)	—	—	—	—	—	—	—	1 (1)
AGAGG	—	1 (1)	—	—	—	—	—	—	—	1 (1)	—	—	—
AGCGC	—	—	—	—	—	—	—	—	—	—	—	—	1 (1)
AGGCG	—	—	—	—	—	—	—	—	—	—	—	—	1 (1)
AAAAAG	—	—	—	2 (2)	—	4 (3)	3 (3)	—	—	—	—	—	—
AAAATC	—	—	—	—	1 (1)	—	—	—	—	—	—	—	—
AAACAC	—	—	—	1 (1)	6 (4)	1 (1)	6 (5)	—	—	—	—	—	1 (1)
AAACCC	1 (1)	—	—	—	—	—	—	—	—	—	—	—	—
AAACCT	1 (1)	—	—	—	—	—	—	—	—	—	—	—	—
AAACGC	—	—	—	—	2 (2)	—	—	—	—	—	—	—	—
AAAGAG	1 (1)	3 (2)	—	1 (1)	2 (1)	—	—	—	—	—	—	—	—
AAAGAT	1 (1)	—	—	—	—	—	—	—	—	—	—	—	—
AAAGCC	—	—	—	1 (1)	5 (5)	—	—	—	—	—	—	—	—
AAAGGT	—	—	—	—	—	—	—	—	—	—	—	—	2 (2)
AAATTC	—	1 (1)	—	—	—	—	—	—	—	—	—	—	—
AACAAG	—	—	—	—	—	—	—	—	—	—	—	—	1 (1)
AACACC	—	—	—	—	—	—	—	—	—	—	—	—	1 (1)
AACAGC	—	—	—	—	—	—	—	—	—	—	—	—	2 (2)
AACATC	—	—	—	—	—	—	—	—	—	—	—	—	1 (1)
AACCCT	1 (1)	—	—	—	—	—	—	—	—	—	—	—	—
AACGAC	—	—	—	3 (2)	—	—	2 (2)	—	—	—	—	—	—
AAGACC	—	—	—	—	—	—	—	—	—	—	—	—	1 (1)
AAGAGC	—	—	—	—	—	—	—	—	—	—	—	—	4 (2)
AAGAGG	—	—	—	1 (1)	—	—	—	—	1 (1)	1 (1)	—	—	—

continued

Table 3. *Continued*

AAGCAC	—	—	—	—	—	—	—	—	—	—	—	—	1 (1)
AAGCAT	—	—	—	—	—	—	—	—	—	—	—	—	2 (1)
AAGCCC	—	—	—	1 (1)	5 (5)	—	—	—	—	—	—	—	—
AAGGAG	—	—	—	—	1 (1)	2 (1)	3 (2)	—	—	—	—	—	6 (3)
AAGGCC	—	—	—	—	—	—	—	—	—	—	—	—	2 (2)
AAGTGT	—	—	—	—	—	6 (3)	3 (2)	—	—	—	—	—	—
AATACC	—	—	—	—	—	—	—	—	—	—	—	—	1 (1)
AATCCG	1 (1)	—	—	—	—	—	—	—	—	—	—	—	—
ACACAG	1 (1)	—	—	—	—	—	—	—	—	—	—	—	7 (7)
ACACAT	—	—	2 (2)	5 (5)	1 (1)	—	1 (1)	—	1 (1)	—	—	—	1 (1)
ACACCC	—	—	—	1 (1)	—	—	—	—	—	—	—	—	1 (1)
ACACCT	—	—	—	—	—	—	1 (4)	—	—	—	—	—	—
ACACGC	2 (2)	2 (2)	—	3 (3)	7 (7)	6 (3)	3 (2)	—	—	—	—	—	16 (16)
ACACGT	—	—	—	—	—	—	—	—	—	—	—	—	1 (1)
ACACTC	—	—	—	—	1 (1)	—	—	—	—	—	—	—	1 (1)
ACACTG	—	—	—	—	—	—	—	—	—	—	—	—	1 (1)
ACAGAG	—	—	1 (1)	4 (4)	—	—	—	—	—	—	—	—	5 (5)
ACAGAT	—	—	—	—	—	—	—	—	—	—	—	—	1 (1)
ACAGCC	—	—	—	—	—	—	—	—	—	—	—	—	1 (1)
ACAGGT	—	—	—	1 (1)	—	—	—	—	—	—	—	—	—
ACATAG	—	—	—	—	—	—	—	—	—	—	—	—	1 (1)
ACATAT	—	—	—	—	—	1 (1)	1 (1)	—	1 (1)	—	—	—	1 (1)
ACATCT	—	—	—	—	—	—	1 (1)	—	—	—	—	—	—
ACATGG	—	—	—	—	—	—	—	—	—	—	—	—	1 (1)
ACCACT	—	—	—	—	—	—	—	—	—	—	—	—	1 (1)
ACCATC	—	—	—	—	—	—	—	—	—	—	—	—	2 (2)
ACCCGC	—	—	—	—	—	—	—	—	—	—	—	—	3 (3)
ACCCTC	—	—	—	—	—	—	—	—	—	—	—	—	1 (1)
ACCGCC	—	—	—	—	—	—	—	—	—	—	—	—	1 (1)
ACCTGG	—	—	—	—	—	—	—	—	—	—	—	—	1 (1)
ACGCAG	—	—	—	—	—	—	—	—	—	—	—	—	1 (1)
ACGCCC	—	—	—	1 (1)	—	—	—	—	—	—	—	—	—
ACGTGC	—	—	—	—	—	—	—	—	—	—	—	—	1 (1)
ACTCCC	—	—	—	—	—	—	—	—	—	—	—	—	1 (1)
ACTGCC	—	2 (2)	—	—	—	—	—	—	—	—	—	—	—
AGAGAT	1 (1)	—	1 (1)	3 (3)	—	—	—	—	1 (1)	—	—	—	—
AGAGGC	—	—	—	—	—	—	—	—	—	—	—	—	2 (2)
AGAGGG	2 (1)	—	—	—	—	—	—	—	—	—	—	—	—
AGCAGG	—	—	—	—	—	—	—	—	—	—	—	—	3 (2)
AGGCAT	—	—	—	—	—	—	—	—	—	—	—	—	3 (1)
ATCGCC	—	—	—	—	—	—	—	—	—	—	—	—	1 (1)

*Target genomes of Ctr = *Ceratoserolis trilobitoides*, Spa = *Serolis paradoxa*, Gly = *Glyptonotus antarcticus*, Bop = Bopyridae, Sse = *Septemserolis septemcarinata*, Mgr = *Munida gregaria*, Pmu = *Pseudo-nitzschia multistriata*, Ata = *Alexandrium tamutum*. Reporter genomes of M = *Mus musculus*, D = *Drosophila melanogaster*, H = *Homo sapiens*.

recovered DNA fragments that were obtained from the same target genome but enriched with different reporter genomes. We compared (Table 2) the library pairs (1,2), (4,5), (6,7), and (8,9). It can be observed (Fig. 4) that, as expected, in each library pair, the relative density of dinucleotide repeats is always highest in the library produced with *M. musculus*. The probabilities (*p* values) that in libraries 1, 4, 6, and 8 the higher density of dinucleotide repeats could have arisen by chance are, respectively, 0.02%, 0.002%, 1.6%, 1.5%, which is highly significant. Similarly, in the library pairs (1,2), (4,5), (6,7), the relative densities of trinucleotide repeats are higher in the libraries created using *D. melanogaster* as reporter genome. The probabilities that an even higher relative trinucleotide density could have arisen by chance are 21%, 0.6%, 5.8%, showing a significant level only for pair (4,5) with *S. paradoxa* as target species.

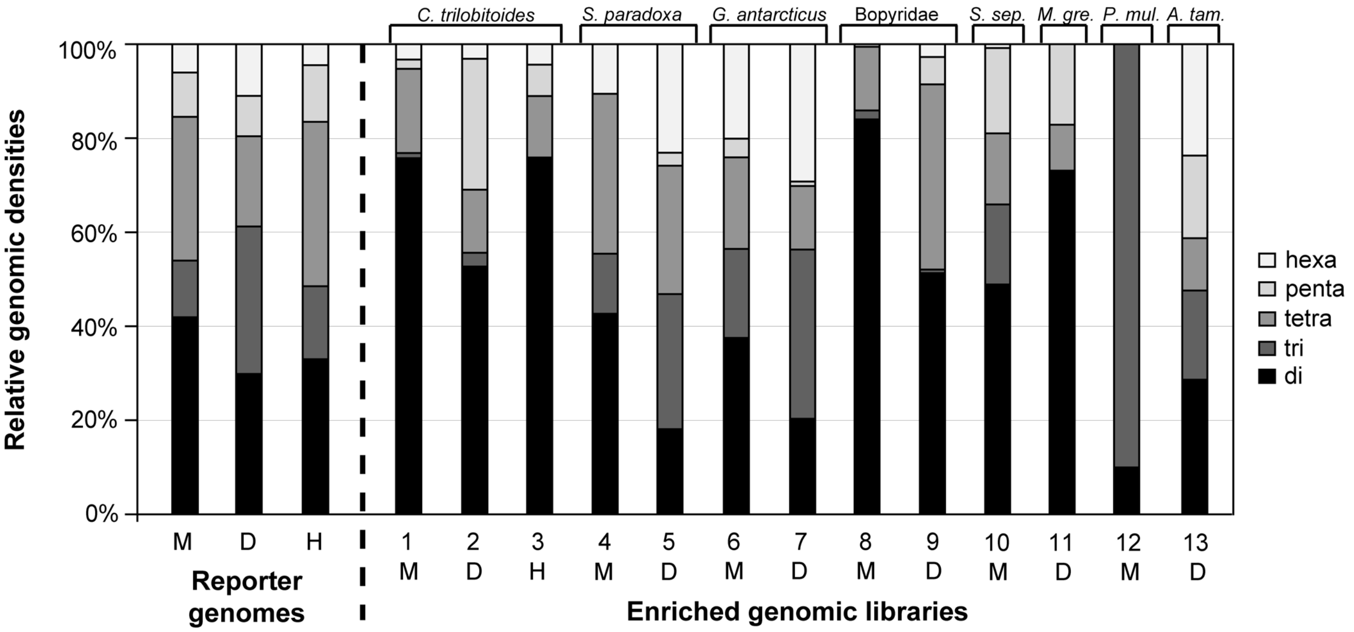

Fig. 4. Relative genomic densities of di- to hexanucleotide repeats in the reporter genomes (M = *Mus musculus*, D = *Drosophila melanogaster*, H = *Homo sapiens*) as well as relative densities in the 13 enriched genomic libraries. The letters below the library numbers indicate the reporter genome used for enrichment. Several different repeat classes were found in all but library 12 (*Pseudo-nitzschia multistriata*).

Furthermore, it has been tested, whether the repeat type densities of $(AC)_n$, $(AG)_n$, and $(AT)_n$ repeats in the reporter genome had an influence on the enriched libraries (Fig. 5).

The relative densities of $(AC)_n$ repeats were significantly higher for the set of genomic libraries enriched using *D. melanogaster* as reporter genome compared to using *M. musculus* (*p* values of permutation test for library pairs (1,2): <0.001% (4,5): <0.001% (6,7): 2.6% (8,9): 0.036%). For $(AG)_n$ repeats, the situation is vice versa with libraries enriched using *M. musculus* as reporter genome being significantly more enriched for $(AG)_n$ repeats (*p* values of permutation test for library pairs (1,2): <0.001% (4,5): 0.003% (6,7): 1.7% (8,9): 0.04%). No significant difference for the density of $(AT)_n$ repeats was observed, neither in the reporter genomes nor in the pairs of libraries belonging to the same target species.

Discussion

In exploratory studies based on microsatellite data the use of an efficient enrichment protocol and strategies for subsequent data analysis are essential to obtain many appropriate microsatellite loci within a short time and at low cost. This study showed that the reporter genome protocol is a universally applicable and very efficient enrichment protocol that has unique advantages.

Table 4. Microsatellite characteristis of the reporter genomes*

Reporter genomes	*M. musculus* (2590 Mbp)[†]			*D. melanogaster* (122 Mbp)[‡]			*H. sapiens* (2940 Mbp)[§]		
	Density [bp/Mbp]	Relative amount	Length ∅ ± SD	Density [bp/Mbp]	Relative amount	Length ∅ ± SD	Density [bp/Mbp]	Relative amount	Length ∅ ± SD
Dinucleotide	10044	41.9	28.1 ± 24.5	3869	29.9	15.8 ± 8.1	3302	33.0	18.2 ± 12.2
Trinucleotide	2891	12.1	20.7 ± 26.7	4061	31.3	14.8 ± 8.3	1554	15.5	14.9 ± 9.9
Tetranucleotide	7326	30.6	24.3 ± 25.5	2494	19.2	15.1 ± 7.9	3500	35.0	19.8 ± 16.1
Pentanucleotide	2261	9.4	29.8 ± 30.2	1110	8.6	19.3 ± 7.6	1204	12.0	22.8 ± 14.8
Hexanucleotide	1447	6.0	38.7 ± 44.0	1425	6.0	24.3 ± 28.5	449	4.5	25.2 ± 18.2

*The table heading contains the total count, number of different repeat types, and density of microsatellites in the three genomes. The table rows show the densities, relative amounts, and length characteristics of the five repeat classes. Phobos search parameters are given in the *Material and procedures* section.

[†] 2424254 microsatellites, 476 different repeat types, density 23733 bp/Mbp

[‡] 96779 microsatellites, 467 different repeat types, density 12924 bp/Mbp

[§] 1651071 microsatellites, 454 different repeat types, density 9927 bp/Mbp

Efficiency of the reporter genome protocol—The average success rate of the reporter genome protocol within the 13 genomic libraries created from eight taxa was 85.5% (90.5% without the outlier *S. septemcarinata*). This is comparable to the success of other recent protocols (Glenn and Schable 2005; Korpelainen et al. 2007; Zane et al. 2002). In our laboratory, previous attempts to isolate microsatellites from *C. trilobitoides* using the screening of (i) a nonenriched library and (ii) enriched libraries with radioactively labeled, synthetic oligomer probes (Rassmann et al. 1991), and (iii) the identification of microsatellites using the RAPD-based PIMA protocol (Lunt et al. 1999) were conducted. Whereas protocol (i) and (ii) failed to isolate sufficiently many appropriate microsatellites in an acceptable time, protocol (iii) failed entirely.

The lower success rate of the enrichment using *S. septemcarinata* as target genome (25%) was still high enough to identify 63 microsatellites within the 146 clones. This was sufficient to establish an informative set of polymorphic markers for population genetic studies (Leese et al. 2008).

This study also indicates that there is no significant difference in overall enrichment efficiency regarding the three different taxonomic target groups. Consequently, the reporter genome protocol is capable of detecting microsatellites from many different eukaryotic genomes without prior knowledge about repeat classes and their frequencies in the target genome.

To estimate an enrichment factor for a given repeat type, its frequency in the genome has to be known. For *C. trilobitoides*, screening a nonenriched library for $(AC)_n$ repeats allowed the calculation of the genomic density of $(AC)_n$ repeats to be approximately 343 bp/Mbp (see Assessment section). Cloning efficiency for $(AC)_n$ repeats was 26/3456 = 0.75%, which means that every 134th clone contains a $(AC)_n$ repeat on average. In the three enriched libraries for *C. trilobitoides* (libraries 1-3, Table 3), the proportion of clones with at least one $(AC)_n$ repeat was 25%, 43%, and 60% for *M. musculus*, *H. sapiens,* and *D. melanogaster* as reporter genomes corresponding to enrichment factors of 33.3, 57.3, and 80.0 for this repeat type. Alternatively, enrichment factors can be defined as the ratio of repeat type densities in enriched versus nonenriched partial genomic libraries. For $(AC)_n$ repeats in libraries 1-3, this yields enrichment factors as high as 104.7, 212.8, and 180.2. Even though the definition of enrichment factors based on repeat densities seems straightforward, it is biased if multiple (i.e., hitchhiking) microsatellites occur in the same insert. The two different approaches for defining an enrichment factor yielded considerably different results reflecting the basic difficulty in defining an appropriate and meaningful measure for an enrichment of microsatellites.

The density of $(AC)_n$ repeats found in the nonenriched partial genomic library of *C. trilobitoides* in this study can be compared with densities found in other genomes analyzed. The average genomic density of $(AC)_n$ repeats in arthropods estimated by Tóth et al. (2000) is 825 bp/Mbp for perfect repeats longer than 12 bp, which is much higher than the 343 bp/Mbp for imperfect repeats longer than 9 bp in *C. trilobitoides*. For a comparison with identical search parameters, we computed the genomic densities of $(AC)_n$ repeats for the three

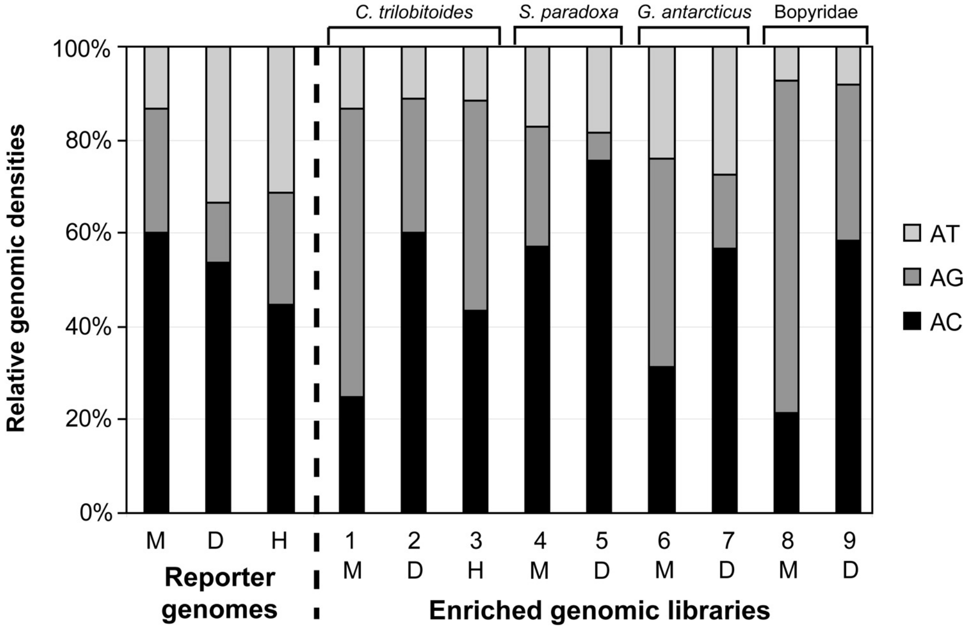

Fig. 5. Relative genomic densities of the three major dinucleotide repeats $(AC)_n$, $(AG)_n$, and $(AT)_n$ in the reporter genomes (M = *Mus musculus*, D = *Drosophila melanogaster*, and H = *Homo sapiens*) as well as in the enriched genomic libraries of the four isopods *Ceratoserolis trilobitoides* (1-3), *Serolis paradoxa* (4,5), *Glyptonotus antarcticus* (6,7), and the bopyrid isopods (8,9). The letters below the library numbers indicate the reporter genome used for enrichment.

fully sequenced arthropods *Drosophila melanogaster*, *Apis mellifera,* and *Daphnia pulex* which are 2138, 583, and 1387 bp/Mbp*,* respectively (unpubl. data). Consequently, *C. trilobitoides* seems to have a low density of $(AC)_n$ repeats for an arthropod species.

Influence of the reporter genomes—The three reporter genomes used in this study differed considerably in their global microsatellite densities, their length characteristics, and in the relative frequencies of different repeat classes (Table 4, Fig. 4).

As a first result, the global microsatellite density of the reporter genome does not seem to have a significant influence on protocol efficiency: whereas *D. melanogaster* and *H. sapiens* have a microsatellite density of about half of that of *M. musculus* (Table 4), protocol performance was equally high for *M. musculus* and *H. sapiens* (libraries 1 and 3), but was significantly lower for *D. melanogaster* as reporter genome in the library pairs (1,2), (4,5), (6,7), and (8,9). This indicates that the number of hybridization sites is sufficient in all three reporter genomes used in this study and that protocol performance is independent of the microsatellite density of reporter genomes in this density range (Table 2). In the present data set, protocol efficiency is best explained with genetic distances between target and reporter genome. The smaller the genetic distance between target and reporter species, the more hybridizations occur at loci other than microsatellites, which lowers the enrichment. The best example is library 2, where a BLAST search revealed 19 target genome inserts of *C. trilobitoides* to be highly similar to conserved regions of the 18S and 28S rDNA regions of the reporter genome (*D. melanogaster*). This was avoided when using DNA of *M. musculus* or *H. sapiens* as reporter genomes indicating that sequence similarity between the two arthropod species *C. trilobitoides* and *D. melanogaster* was still high enough to allow hybridization of homologous loci, thus defying the purpose of the enrichment procedure.

The relative repeat class densities of di- and trinucleotide repeats in the reporter genomes have an influence on the relative repeat class densities in the enriched library. A permutation test confirmed that the considerably higher relative dinucleotide repeat density of *M. musculus* as compared with *D. melanogaster* lead to a significantly higher enrichment of this repeat class in all four library pairs with the same target species and these two reporter genomes. The higher relative trinucleotide repeat density of *D. melanogaster* as compared with *M. musculus* lead to a significant effect only in the library pair (4,5), whereas in the other library pairs the data basis is too small to get significant results. Tetra- to hexanucleotide repeat densities were not compared, since their repeat densities were usually high only in very few inserts of the partial genomic libraries, making a meaningful statistical comparison impossible.

The influence of the reporter genome can even be traced down to individual repeat types. It has been shown that the libraries produced with *M. musculus* as reporter genome contain significantly more $(AG)_n$ repeats, whereas the libraries produced with *D. melanogaster* contain significantly more $(AC)_n$ repeats.

It is difficult to infer exactly how strongly the repeat content of a reporter genome influences the libraries produced with it and whether it is possible to even find an approximate functional relationship. There are at least four main limiting factors here: (i) hitchhiking repeats which are only in the library due to their occurrence in the same insert as a repeat that has hybridized to the reporter genome, (ii) biochemical differences among repeat types, such as different binding strengths, (iii) the ability of some repeat types to form secondary structures, and (iv) different mean imperfections of repeat types. These limiting factors might make it impossible to find an exact functional relationship. Ignoring these limiting factors, the probability that a certain repeat type in the target genome hybridizes against the reporter is proportional to the repeat density in each of the two genomes. Thus the expected density of this repeat type in the library is

$$D_{\text{library}}(\text{repeat type}) \sim D_{\text{reporter}}(\text{repeat type}) * D_{\text{target}}(\text{repeat type}).$$

It was tested whether this relationship holds by means of the ratio of $(AC)_n$ to $(AG)_n$ repeat densities in all libraries. However, this relationship could not be confirmed, since *D. melanogaster* reports much more $(AC)_n$ repeats and *M. musculus* reports much more $(AG)_n$ repeats than expected from this relationship.

Information on the dominant repeat types in target genome—For the genomic libraries created in this study, we showed that the repeat content of the reporter genome has a significant influence on the densities in the libraries, even though a direct proportionality relation had to be rejected. This makes it possible to retain at least some information about the dominant repeat types in the target genome. In the enriched libraries (1-3, Table 3, Fig. 5) of *C. trilobitoides* the AG/AC and AG/AT density ratios are significantly higher than in the corresponding reporter genomes. The same holds true for libraries (6-9), indicating that for *C. trilobitoides*, *G. antarcticus*, and the bopyrid isopod, the $(AG)_n$ repeat type is likely to be a dominant repeat type. For the two libraries of *S. paradoxa*, however, the results are ambiguous.

From a comparative genomics point of view, a high $(AG)_n$ repeat content has been reported from only few species (e.g., Estoup et al. 1993; Thoren et al. 1995; Tóth 2000; Tóth et al. 2000; Xu et al. 1999). Generally, $(AC)_n$ and $(AT)_n$ dinucleotide repeat types constitute the major fraction of repeats in other crustaceans investigated so far (Katti et al. 2001; Kong and Gao 2005, Mayer et al. in prep.; Tassanakajon et al. 1998).

Reduced selection bias of the reporter genome protocol—The reporter genome protocol substitutes synthetic repeat probes with repeats naturally occurring in whole genomes as hybridization templates (Fig. 2). Using the repeat detection and analysis tools Phobos and Sat-Stat, we have identified 476 different and independent di- to hexanucleotide repeat types in the reporter genome of *M. musculus*, 454 repeat types in *H. sapiens*, and 467 repeat types in *D. melanogaster* (Table 4). Combining the three reporter genomes, all 499 possible different and independent repeat types are available as hybridization probes. Although many motifs are rare in the reporter genomes, this still ensures a much less biased search than standard protocols which rarely screen for more than 20 different repeat types. Thus, without knowledge about relative densities of certain repeat types, which are commonly dissimilar among different taxa (Tóth et al. 2000), the reporter genome protocol allows the isolation of many different microsatellites that are frequent in the target genome without prior knowledge about repeat motifs (see Table 2). This significantly reduces the selection bias for the reporter genome protocol compared to all other protocols and explains its comparatively high success in detecting as many as 133 different repeat types in the 13 genomic libraries screened in this study.

Comments and recommendations

Resources and hand-on time—Overall, resource-usage and handling time of the RGP is comparable to the selective hybridization protocols discussed in Zane et al. (2002). The RGP requires a standard molecular biology laboratory with cloning facilities. There is no need, e.g., for isotopic or specialized hybridization equipment. Its major advantage is that it does not rely on the use of specific repeat probes, neither for enriching the fragments for the genomic library nor for screening the genomic library. An enriched genomic library can be established within 1 week. Due to the high success rate, it is generally sufficient to shot-gun sequence one to a few 96-well plates of clones to obtain sufficiently many informative microsatellite loci for standard exploratory studies.

Choice of reporter genomes—Our study suggests choosing reporter genomes distantly related to the target genomes to avoid hybridization at homologous nonrepetitive regions. Using completely sequenced reporter genomes has the advantage that rare but possible contaminations by mobilized reporter genome fragments can be detected by BLAST searches. If a great diversity of different repeat types is desired, it is useful to select at least two reporter genomes that differ in their relative genomic repeat type densities, e.g., *Mus musculus* and *Drosophila melanogaster*.

Microsatellites for population genetic studies—For population genetic studies, highly perfect microsatellites (Fig. 3b,c) are preferred over imperfect, compound, or complex microsatellites (Fig. 3d-f). The latter are more likely to have not just one but several sources of length variation and therefore show a higher degree of size homoplasy that can only be reliably assessed by sequencing (Adams et al. 1993; Estoup et al. 1995, 2002). For population genetic and pedigree analyses, longer microsatellites (>8 repeat units) are preferred over shorter microsatellites as their higher variability allows a better resolution. Very long microsatellites often suffer from in vitro artefacts (Shinde et al. 2003), and are frequently biased in mutation toward shorter repeats (Ellegren 2004; Nauta and Weissing 1996; Wierdl et al. 1997; Xu et al. 2000). This advises using microsatellite markers with <40 repeat units for analysis, especially for dinucleotide repeats. Besides the structure of a microsatellite itself, a crucial point is the quality of its flanking regions. Frequently, additional structural processes, e.g., inversions, duplications, or higher order repeat structures (see Fig. 3 g-f), can be detected in the flanking regions, which may complicate or bias analysis or even lead to the complete failure of PCR amplification. The process of identifying suitable markers from the large amount of microsatellites was greatly facilitated by a software workflow based on the Staden software package, a version of which will soon be available (Kraemer et al. in prep.) including Phobos and other modules, geared toward high throughput establishment of high quality microsatellite loci.

A common feature of enriched libraries is that many microsatellite loci isolated from them have disadvantageous properties and should be excluded for various reasons (Fig. 1). However, the number of microsatellites found in the enriched libraries in this study was so high that a sufficient number of appropriate, high quality loci were available for primer design (see Leese and Held (2008) and Leese et al. (2008) for details on primer design).

Conclusions

The reporter genome protocol is a highly efficient protocol to isolate microsatellites from unknown genomes. Its major advantages are (i) the high enrichment efficiency across a large taxonomic diversity of target genomes, (ii) the large number of different repeat types that can be detected without the need to specify synthetic reporter probes, and (iii) the possibility to retain some information about relative densities of repeat types in the target genome.

The hybridization step is expected to yield maximum numbers of microsatellites if the target and the reporter genomes share similarities in repeat regions only, but are sufficiently dissimilar by descent outside repeat regions. We are confident that the reporter genome protocol as outlined in this study will make microsatellites more attractive for researchers in a wide variety of fields that previously avoided the development of this powerful marker system.

References

Adams, R. P., T. Demeke, and H. A. Abulfaith. 1993. RAPD DNA fingerprints and terpenoids: Clue to past migrations of Juniperus in Arabia and east Africa. Theor. Appl. Genet. 87:22-26.

Armour, J. A., R. Neumann, S. Gobert, and A. J. Jeffreys. 1994. Isolation of human simple repeat loci by hybridization selection. Hum. Mol. Genet. 3:599-605

Balding, D. 1999. Forensic applications of microsatellite markers, p. 198-210. *In* D. B. Goldstein and C. Schlötterer [eds.], Microsatellites - evolution and applications. Oxford Univ. Press.

Benson, G. 1999. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. 27:573-580.

Bowcock, A. M., A. Ruiz-Linares, J. Tomfohrde, E. Minch, J. R. Kidd, and L. L. Cavalli-Sforza. 1994. High resolution of human evolutionary trees with polymorphic microsatellites. Nature 368:455-457.

Chambers, G. K., and E. S. MacAvoy. 2000. Microsatellites: consensus and controversy. Comp. Biochem. Physiol. B Biochem. Mol. Biol. 126:455-476.

Dib, C., and others. 1996. A comprehensive genetic map of the human genome based on 5264 microsatellites. Nature 380:152-154.

Ellegren, H. 2004. Microsatellites: Simple sequences with complex evolution. Nat. Rev. Genet. 5:435-445.

Estoup, A., P. Jarne, and J. M. Cornuet. 2002. Homoplasy and mutation model at microsatellite loci and their consequences for population genetic analysis. Mol. Ecol. 11:1591-1604.

———, M. Solignac, M. Harry, and J. M. Cornuet. 1993. Characterization of $(GT)_n$ and $(CT)_n$ microsatellites in two insect species: *Apis mellifera* and *Bombus terrestris*. Nucleic Acids Res. 21:1427-1431.

———, C. Tailliez, J. M. Cornuet, and M. Solignac. 1995. Size homoplasy and mutational processes of interrupted microsatellites in two bee species, *Apis mellifera* and *Bombus terrestris* (Apidae). Mol. Biol. Evol. 12:1074-1084.

Ewing, B., and P. Green. 1998. Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res. 8:186-194.

———, L. Hillier, M. C. Wendl, and P. Green. 1998. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res. 8:175-185.

Glenn, T. C., and N. A. Schable. 2005. Isolating microsatellite DNA loci. Methods Enzymol. 395:202-222.

Held, C. 2003. Molecular evidence for cryptic speciation within the widespread Antarctic crustacean *Ceratoserolis trilobitoides* (Crustacea, Isopoda), p. 135-139. *In* A. H. L. Huiskes et al. [eds.], Antarctic biology in a global context. Backhuys Publishers.

Jones, M. W. and others. 2003. Development, characterisation, inheritance, and cross-species utility of American lobster (*Homarus americanus*) microsatellite and mtDNA PCR-RFLP markers. Genome 46:59-69.

Junier, T., and M. Pagni. 2000. Dotlet: diagonal plots in a web browser. Bioinformatics 16:178-179.

Kandpal, R. P., G. Kandpal, and S. M. Weissman. 1994. Construction of libraries enriched for sequence repeats and jumping clones, and hybridization selection for region-specific markers. Proc. Nat. Acad. Sci. USA 91:88-92.

Karagyozov, L., I. D. Kalcheva, and V. M. Chapman. 1993. Construction of random small-insert genomic libraries highly enriched for simple sequence repeats. Nucleic Acids Res. 21:3911-3912.

Katti, M. V., P. K. Ranjekar, and V. S. Gupta. 2001. Differential distribution of simple sequence repeats in eukaryotic genome sequences. Mol. Biol. Evol. 18:1161-1167.

Kijas, J. M., J. C. Fowler, C. A. Garbett, and M. R. Thomas. 1994. Enrichment of microsatellites from the citrus genome using biotinylated oligonucleotide sequences bound to streptavidin-coated magnetic beads. BioTechniques 16:656-662.

Kong, J., and H. A. Gao. 2005. Analysis of tandem repeats in the genome of Chinese shrimp *Fenneropenaeus chinensis*. Chin. Sci. Bull. 50:1462-1469.

Korpelainen, H., K. Kostamo, and V. Virtanen. 2007. Microsatellite marker identification using genome screening and restriction-ligation. BioTechniques 42:479-486.

Leese, F., and C. Held. 2008. Identification and characterization of microsatellites from the Antarctic isopod *Ceratoserolis trilobitoides*—nuclear evidence for cryptic species. Conserv. Genet. doi: 10.1007/s10592-007-9491-z

———, A. Kop, S. Agrawal, and C. Held. 2008. Isolation and characterization of microsatellite markers from the marine serolid isopods *Serolis paradoxa* and *Septemserolis septemcarinata* (Crustacea: Peracarida). Mol. Ecol. Res. doi:10.1111/j.1755-0998.2007.02078.x

Levinson, G., J. L. Marsh, J. T. Epplen, and G. A. Gutman. 1985. Cross-hybridizing snake satellite, *Drosophila*, and mouse DNA sequences may have arisen independently. Mol. Biol. Evol. 2:494-504.

Lunt, D. H., W. F. Hutchinson, and G. R. Carvalho. 1999. An efficient method for PCR-based isolation of microsatellite arrays (PIMA). Mol. Ecol. 8:891-893.

Nauta, M. J., and F. J. Weissing. 1996. Constraints on allele size at microsatellite loci: implications for genetic differentiation. Genetics 143:1021-1032.

Nolte, A. W., K. C. Stemshorn, and D. Tautz. 2005. Direct cloning of microsatellite loci from *Cottus gobio* through a simplified enrichment procedure. Mol. Ecol. Notes 5:628-636.

Paetkau, D., W. Calvert, I. Stirling, and C. Strobeck. 1995. Microsatellite analysis of population-structure in canadian Polar Bears. Mol. Ecol. 4:347-354.

Rassmann, K., C. Schlötterer, and D. Tautz. 1991. Isolation of simple sequence loci for use in polymerase chain reaction based fingerprinting. Electrophoresis 12:113-118.

Rozen, S., and H. J. Skaletsky. 2000. Primer3 on the WWW for general users and for biologist programmers, p. 365-386. *In* S. Krawetz and S. Misener [eds.], Bioinformatics methods and protocols: Methods in molecular biology. Humana Press.

Sambrook, J., E. F. Fritsch, and T. Maniatis. 1989. Molecular cloning. A laboratory manual. 2nd edition. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.

Schlötterer, C. 2004. The evolution of molecular markers - just a matter of fashion. Nat. Rev. Genet. 5:63-69.

Selkoe, K., and R. J. Toonen. 2006. Microsatellites for ecologists: A practical guide to using and evaluating microsatellite markers. Ecol. Lett. 9:615-629.

Shinde, D., Y. L. Lai, F. Z. Sun, and N. Arnheim. 2003. Taq DNA polymerase slippage mutation rates measured by PCR and quasi-likelihood analysis: (CA/GT)(n) and (A/T)(n) microsatellites. Nucleic Acids Res. 31:974-980.

Staden, R. 1996. The Staden sequence analysis package. Mol. Biotechnol. 5:233-241.

Tassanakajon, A., and others. 1998. Isolation and characterization of microsatellite markers in the black tiger prawn *Penaeus monodon*. Mol. Mar. Biol. Biotechnol. 7:55-61.

Thoren, P. A., R. J. Paxton, and A. Estoup. 1995. Unusually high frequency of $(CT)_n$ and $(GT)_n$ microsatellite loci in a yellowjacket wasp, *Vespula rufa* (L.) (Hymenoptera: Vespidae). Insect. Mol. Biol 4:141-148.

Tóth, G. 2000. SSRDB: A database of simple sequence repeats (SSRs) in eukaryotes. http://bioinformatics.abc.hu/ssr/ (online database).

———, Z. Gaspari, and J. Jurka. 2000. Microsatellites in different eukaryotic genomes: Survey and analysis. Genome Res. 10:967-981.

Vos, P., and others. 1995. AFLP: a new technique for DNA fingerprinting. Nucleic Acids Res. 23:4407-4414.

Weetman, D., A. Ruggiero, S. Mariani, P. W. Shaw, A. R. Lawler, and L. Hauser. 2007. Hierarchical population genetic structure in the commercially exploited shrimp *Crangon crangon* identified by AFLP analysis. Mar. Biol. (Berl) 151:565-575.

Weissenbach, J., and others. 1992. A second-generation linkage map of the human genome. Nature 359:794-801.

Wierdl, M., M. Dominska, and T. D. Petes. 1997. Microsatellite instability in yeast: Dependence of the length of the microsatellite. Genetics 146:769-779.

Xu, A. K. Dhar, J. Wyrzykowski, and A. Alcivar-Warren. 1999. Identification of abundant and informative microsatellites from shrimp (*Penaeus monodon*) genome. Anim. Genet. 30:150-156.

———, M. Peng, Z. Fang, and X. P. Xu. 2000. The direction of microsatellite mutations is dependent upon allele length. Nature Gen. 24:396-399.

Zane, L., L. Bargelloni, and T. Patarnello. 2002. Strategies for microsatellite isolation: a review. Mol. Ecol. 11:1-16.

Zeller, M., T. B. H. Reusch, and W. Lampert. 2006. A comparative population genetic study on calanoid freshwater copepods: Investigation of isolation-by-distance in two Eudiaptomus species with a different potential for dispersal. Limnol. Oceanogr. 51:117-124.

Submitted 10 December 2007
Revised 16 May 2008
Accepted 27 June 2008