High-throughput microsatellite isolation through 454 GS-FLX Titanium pyrosequencing of enriched DNA libraries

Thibaut Malausa, André Gilles, Emese Meglécz, Hélène Blancard, Stéphanie Duthoy, Caroline Costedoat, Vincent Dubut, Nicolas Pech, Philippe Castagnone-Sereno, Christophe Delye, et al.

▶ **To cite this version:**

Thibaut Malausa, André Gilles, Emese Meglécz, Hélène Blancard, Stéphanie Duthoy, et al.. High-throughput microsatellite isolation through 454 GS-FLX Titanium pyrosequencing of enriched DNA libraries. Molecular Ecology Resources, 2011, 11 (4), pp.638-644. 10.1111/j.1755-0998.2011.02992.x . hal-01506135

HAL Id: hal-01506135
https://hal.science/hal-01506135

Submitted on 16 Sep 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

High-throughput microsatellite isolation through 454 GS-FLX Titanium pyrosequencing of enriched DNA libraries

THIBAUT MALAUSA,* ANDRÉ GILLES,† EMESE MEGLÉCZ,† HÉLÈNE BLANQUART,‡ STÉPHANIE DUTHOY,‡ CAROLINE COSTEDOAT,† VINCENT DUBUT,† NICOLAS PECH,† PHILIPPE CASTAGNONE-SERENO,* CHRISTOPHE DÉLYE,§ NICOLAS FEAU,¶ PASCAL FREY,** PHILIPPE GAUTHIER,†† THOMAS GUILLEMAUD,* LAURENT HAZARD,‡‡ VALÉRIE LE CORRE,§ BRIGITTE LUNG-ESCARMANT,¶ PIERRE-JEAN G. MALÉ,§§ STÉPHANIE FERREIRA‡ and JEAN-FRANÇOIS MARTIN††

**INRA, UMR 1301 IBSV INRA/UNSA/CNRS, 400 Route des Chappes, BP 167, 06903 Sophia-Antipolis Cedex, France, †Aix-Marseille Université, CNRS, IRD, UMR 6116 IMEP, Equipe Evolution Génome Environnement, Centre Saint-Charles, Case 36, 3 Place Victor Hugo, 13331 Marseille Cedex 3, France, ‡Genoscreen, Genomic Platform and R&D, Campus de l'Institut Pasteur, 1 rue du Professeur Calmette, Bâtiment Guérin, 59000 Lille, France, §INRA, UMR 1210 Biologie et Gestion des Adventices, 17 rue Sully, 21000 Dijon, France, ¶INRA, UMR 1202 BIOGECO, Equipe de Pathologie Forestière, Domaine de Pierroton, 69 route d'Arcachon, 33612 Cestas Cedex, France, **INRA, Nancy-Université, UMR 1136, Interactions Arbres Microorganismes, IFR 110, 54280 Champenoux, France, ††UMR CBGP (INRA/IRD/Cirad/Montpellier SupAgro), Campus International de Baillarguet, CS 30016, 34988 Montferrier-sur-Lez Cedex, France, ‡‡INRA UMR 1248 AGIR, BP 52627, 31326 Castanet-Tolosan Cedex, France, §§UMR Evolution et Diversité Biologique (Université Toulouse III; CNRS), 118 Route de Narbonne, 31062 Toulouse, France*

Abstract

Microsatellites (or SSRs: simple sequence repeats) are among the most frequently used DNA markers in many areas of research. The use of microsatellite markers is limited by the difficulties involved in their *de novo* isolation from species for which no genomic resources are available. We describe here a high-throughput method for isolating microsatellite markers based on coupling multiplex microsatellite enrichment and next-generation sequencing on 454 GS-FLX Titanium platforms. The procedure was calibrated on a model species (*Apis mellifera*) and validated on 13 other species from various taxonomic groups (animals, plants and fungi), including taxa for which severe difficulties were previously encountered using traditional methods. We obtained from 11 497 to 34 483 sequences depending on the species and the number of detected microsatellite loci ranged from 199 to 5791. We thus demonstrated that this procedure can be readily and successfully applied to a large variety of taxonomic groups, at much lower cost than would have been possible with traditional protocols. This method is expected to speed up the acquisition of high-quality genetic markers for nonmodel organisms.

Keywords: enriched library, genetic marker, genotyping, microsatellite isolation, next-generation sequencing, primer design, pyrosequencing

Introduction

Microsatellites (or SSRs: simple sequence repeats) are among the most frequently used DNA markers in many areas of research (Sunnucks 2000). However, their availability and quality are limited by the difficulties of *de novo* development in species for which no genomic information is available. The most commonly used procedure (enrichment of genomic DNA in microsatellite motifs, cloning and sequencing of the enriched DNA library by the Sanger method) is difficult, time-consuming and costly. Enrichment methods generally use a few specific repeated motifs, generally selected without prior knowledge of their abundance in the genome (Castoe *et al.* 2010), hence introducing potential bias in genome representativeness. Furthermore, only a few hundred sequences are generally obtained because of the high cost of Sanger sequencing. Next-generation sequencing through 454 GS-FLX technology (Roche Applied Science)

Correspondence: Thibaut Malausa, Fax: +33 492386401;
E mail: thibaut.malausa@sophia.inra.fr

has opened up new opportunities for microsatellite isolation. First, large shotgun genomic libraries have proved sufficient to isolate a satisfactory number of markers in a few studies (Abdelkrim *et al.* 2009; Allentoft *et al.* 2009; Castoe *et al.* 2010). Second, 454 GS-FLX sequencing can be used to sequence enriched library, thus offering a higher cost-efficiency (Santana *et al.* 2009). However, the use of pyrosequencing applied to microsatellite isolation has remained rare. This situation could evolve quickly if new procedures taking profit of pyrosequencing technology updates (sequencing longer and more numerous DNA fragments) and easily accessible for research teams could be set up. We present here a new method for high-throughput microsatellite isolation combining DNA enrichment procedures with the use of multiplexed microsatellite probes and the update Titanium of the 454 GS-FLX technology. This method was initially developed in the model species *Apis mellifera* (Linnaeus, 1758) [Insecta: Hymenoptera: Apidae], a species with a genome particularly rich in microsatellites (Solignac *et al.* 2007). Its efficiency was subsequently assessed and validated with 13 other species from various taxonomic groups, including fungi, plants and animals. This procedure is now available on our platform (Lille, France) for any research team interested in rapid and low-cost development of wide SSR libraries.

Methods

Construction and sequencing of multiplex-enriched libraries

An optimization of classical biotin-enrichment methods (Kijas *et al.* 1994) was used to prepare the enriched libraries. Genomic DNA was extracted from various tissues (depending on the species), with the DNeasy Tissue Kit (QIAGEN) and the DNeasy Plant Mini Kit (QIAGEN). Enrichment was carried out at Genoscreen (Lille, France), according to the following procedure. Genomic DNA (1 μg) was sonicated or digested with *Rsa*I (FERMENTAS) for 1 h at 37 °C, according to the manufacturer's recommendations, and was then ligated to standard adapters (Adap-F: GTTTAAGGCCTAGCTAGCAGAATC and Adap-R: GATTCTGCTAGCTAGGCCTT). This step was repeated until the average length of DNA fragment was <1500 bp. Samples were then purified on a Nucleofast PCR plate (MACHEREY-NAGEL). Eight biotin-labelled oligonucleotides, corresponding to eight targeted microsatellite motifs, were hybridized to the ligated DNA at 56 °C for 20 min, after initial denaturation of the ligated DNA. The choice of the targeted microsatellite motifs was based on the screening of thirteen published genome sequences or whole-genome shotgun (WGS) sequences (insects: *A. mellifera, Anopheles gambiae, Drosophila melanogaster, D. yakuba, D. simulans, Bombyx mori, Tribolium castaneum*; Vertebrates: *Takifugu rubripes, Danio rerio, Gallus gallus, Bos taurus, Mus musculus* and *Rattus norvegicus*). All perfect microsatellites with at least five repetitions for all di-hexa motifs were extracted. We then identified the 12 most frequent motifs for each genome (Table S1, Supporting information). From this pool of 30 motifs, we selected eight. This selection was based on (i) motif frequencies in different genomes and (ii) melting temperature compatibility (56 °C). At the same time, we avoided using motifs likely to produce hairpin structures, even if highly frequent (e.g. AT, CG and AAT). The following eight probes were designed to enrich total DNA in these motifs: $(AG)_{10}$, $(AC)_{10}$, $(AAC)_8$, $(AGG)_8$, $(ACG)_8$, $(AAG)_8$, $(ACAT)_6$ and $(ATCT)_6$.

The enrichment step was completed with Dynabeads (INVITROGEN). The resulting enriched DNA was amplified with primers corresponding to the library adapters, over 25 cycles (20 s at 95 °C, 20 s at 60 °C and 90 s at 72 °C) and a final extension step of 30 min at 72 °C. The PCR products were purified with a QIAquick PCR purification kit (QIAGEN).

The sample concentration of purified PCR products was determined by quantifying Picogreen fluorescence (Invitrogen), and the fragment size distribution was determined by running 1 μL of each sample on an Agilent Bioanalyzer 2100, using a DNA 7500 chip (Agilent Technologies). The following manufacturer's protocols were carried out at Genoscreen (Lille, France): fragment end polishing, adaptor ligation, during which specific multiplex identifiers (MIDs) were added, library immobilization, fill-in reaction and single-stranded DNA library isolation. The small fragment removal step was not included to avoid the loss of any genetic information. The single-strand DNA profile and quantification were determined by running 1 μL of each sample on an Agilent Bioanalyzer 2100 with a RNA Pico 6000 chip. The concentration (pg/μL) obtained was then used to calculate the number of molecules of the final product/μL: [single-strand DNA (pg/μL)]/[MW of nucleotide (325) × base pair length of DNA strand] × $[6.02 \times 10^{23}]$. The single-strand templates were subsequently diluted to a normalized concentration of 1×10^8 molecules/μL, and multiplexing by equimolar mixture was performed for the analysis of four samples on a 1/8 GsFLX PTP or eight samples on a 1/4 GsFLX PTP. In each GsFLX PTP region, samples were distinguished thanks to their MIDs. Each multiplex library was previously titrated to accurately determine the number of DNA copies per bead required for maximum sequencing quality. Emulsion PCR and sequencing were then carried out according to the GS-FLX protocol, with no modification.

Calibration test

We set up a calibration test, using DNA samples of *A. mellifera*, to assess the minimum number of 454 loading beads required to obtain sequences of satisfactory quality in sufficient amounts. We loaded calibrated quantities of beads (125 000; 75 000; 50 000 and 25 000 beads) onto four regions of a GsFLX plate delimited by a 16-region gasket.

Enriched vs. shotgun library

We evaluated the benefits of DNA enrichment in microsatellite motifs, by comparing the data obtained after sequencing DNA libraries of *A. mellifera* with and without enrichment. This comparison was made with 125 000 loading beads as it provided the maximum number of sequences in a given 1/16 GsFLX plate.

Validation on 13 taxa

Enriched libraries, generated as described earlier, were constructed for 13 additional taxa from various taxonomic groups (Table 1) to assess the general value of the method. The loading of 75 000 beads was performed for each species (one species on a 1/16 GsFLX plate, four species on 1/8 plate or eight species on 1/4 plate using Roche MIDs) as calibration tests revealed that this number provided highest quality and cost-efficiency (see Results).

Data analysis and automated primer design

The QDD pipeline (Meglécz *et al.* 2010) was used to analyse the 454 sequences and design primers for amplification of the detected microsatellite motifs. Sequences were sorted according to their MID (when used), and the MID sequence was subsequently removed. Enrichment adaptors (Adap-F and Adap-R) were then removed from sequences, and sequences with no detected adapter were discarded. Sequences shorter than 80 bp and sequences containing microsatellite motifs shorter than five repeats were discarded. Sequence similarities were detected through an 'all against all' BLAST analysis. Sequences with significant BLAST hits (e-value = 1E-40, microsatellites being soft-masked) but with flanking region identity levels below 90% were discarded to avoid potential intragenomic multicopy sequences. Using BLAST allowed identification of flanking regions with low, but significant similarities. This is a conservative step that aims to eliminate repetitive sequences (e.g. minisatellites and transposable elements), which are unlikely to provide a clear amplification pattern. Sequences displaying only BLAST hits for which pairwise similarity between the complete

Table 1 Name and systematic position of the species used to set and test the procedure; number and length of sequences generated, Accession no. in the NCBI Short Read Archives. Libraries were also submitted to the Dryad Database, doi:10.5061/dryad.8297 (http://dx.doi.org/10.5061/dryad.8297)

Name	Division	Class	Order	Number of sequences	Length		SRA accession
					Mean	Maximum	
Apis mellifera (shotgun)	Arthropoda	Insecta	Hymenoptera	37 870	275	765	SRS150264.1
A. mellifera (enriched 125K)	Arthropoda	Insecta	Hymenoptera	39 473	251	766	SRS150263.1
A. mellifera (enriched 75K)	Arthropoda	Insecta	Hymenoptera	26 428	258	681	SRS150262.1
A. mellifera (enriched 50K)	Arthropoda	Insecta	Hymenoptera	30 041	259	610	SRS150261.1
A. mellifera (enriched 25K)	Arthropoda	Insecta	Hymenoptera	11 571	259	639	SRS150260.1
Venturia canescens	Arthropoda	Insecta	Hymenoptera	21 716	189	539	SRS140293.2
Euphydryas aurinia	Arthropoda	Insecta	Lepidoptera	11 497	184	562	SRS150273.1
Pseudococcus viburni	Arthropoda	Insecta	Hemiptera	29 528	237	611	SRS150265.1
Diabrotica virgifera	Arthropoda	Insecta	Coleoptera	15 207	259	595	SRS140297.2
Bursaphelenchus xylophilus	Nematoda	Secernentea	Aphelenchida	12 286	240	615	SRS140294.2
Barbus meridionalis	Chordata	Actinopterygii	Cypriniformes	13 010	150	477	SRS150271.1
Danio rerio	Chordata	Actinopterygii	Cypriniformes	15 833	193	543	SRS150272.1
Gerbillus nigeriae	Chordata	Mammalia	Rodentia	21 740	153	884	SRS150274.1
Armillaria ostoyae	Basidiomycota	Agaricomycetes	Agaricales	32 488	179	809	SRS150270.1
Phytophthora alni subsp. *uniformis*	Heterokontophyta	Oomycetes	Peronosporales	34 483	209	997	SRS150269.1
Festuca eskia	Magnoliophyta	Liliopsida	Poales	25 577	246	598	SRS150267.1
Hirtella physophora	Magnoliophyta	Eudicotyledones	Malpighiales	34 316	197	849	SRS150266.1
Papaver rhoeas	Magnoliophyta	Eudicotyledones	Ranunculales	13 825	240	570	SRS150268.1

overlapping part of the flanking regions was over 90% were grouped into contigs aligned by ClustalW, and a 2/3 majority rule was used to build a consensus sequence. These consensus sequences substituted the corresponding single reads. Unique sequences (with no BLAST hit according to our criteria) and consensus sequences were used to form a validated set of sequences that was used for further analyses. Primers were designed automatically using the Primer3 algorithm (Rozen & Skaletsky 2000) implemented within QDD. PCR primers were designed only if (i) the target microsatellite had at least five repeats, (ii) the resulting PCR product was between 80 and 500 bp long, (iii) the flanking region contained, at most, a five-base mononucleotide stretch or two repeats of any di-hexa base-pair motif, (iv) the annealing temperature of primers was between 50 and 64 °C, and the difference in annealing temperature between the forward and the reverse primer was <4 °C and (v) the self-complementarities of primers and the complementarities between primers matched the quality criteria used as default parameters in Primer3. We deliberately chose stringent criteria, as the number of microsatellite loci identified through pyrosequencing is large and the most time-consuming and extensive step is the subsequent validation of the designed primers.

Results

Calibration test

The calibration tests revealed strong positive correlation between the numbers of beads and sequences (R^2=0.851, P = 0.015 through permutation test), and although the quality (as inferred by the percentage of sequences that passed the Roche 454 GsFLX Titanium default quality filters) was not homogeneous among the numbers of loading beads (χ^2 = 142.29, P < 0.0001), there was no significant correlation between beads number and quality (R^2=0.053, P > 0.05). Based on this analysis, the loading of 75 000 beads was retained for the final protocol because this number of beads (i) provided sufficient number of sequences (26 428 sequences) with a high-quality index (63.35% of the sequences passed the 454 GsFLX Titanium quality filters) and (ii) allows the sequencing of up to four enriched libraries onto a 1/8 GsFLX plate or eight libraries onto a 1/4 GsFLX plate, using MIDs.

Enriched vs. shotgun library

We obtained 37 870 and 39 473 sequences for the *A. mellifera* libraries without and with enrichment (hereafter referred to as the 'shotgun' and 'enriched' libraries), respectively (Table 1). Short sequences and sequences without microsatellite motifs were discarded, and the remaining sequences were filtered for redundancy and checked for multiple copies in the data set with our bioinformatics pipeline QDD (Meglécz *et al.* 2010). In total, 5157 and 6230 loci containing microsatellites matching the quality criteria implemented in QDD were detected in the shotgun and enriched libraries, respectively. This high number of loci detected in the shotgun library is not surprising, given the high quantities of microsatellites contained in the genome of this organism (Solignac *et al.* 2007). With the shotgun method, 97% of the validated loci corresponded to sequences observed once in the raw data set, whereas the remaining 3% of the validated loci corresponded to consensus sequences (Table 2). With the enrichment method, 22% of the validated loci corresponded to consensus sequences, taking into account variation between sequences (caused by intraspecific polymorphism and polymerase or 454 sequencing errors). QDD was then used to design primer pairs for each validated locus. The primer design was successful in 2045 loci from the shotgun library and in 2200 loci from the enriched library. The properties of the markers designed from the two libraries differed considerably: the percentage of markers displaying microsatellite motifs consisting exclusively of A/T nucleotides was 39% with the shotgun library and only 8% with the enriched library (Table 2). Interestingly, enrichment allowed isolation of around 3 times more primer pairs (543 vs. 186) designed around microsatellite motifs of more than eight perfect repeats (excluding AT-motifs), i.e. optimal microsatellite markers.

Validation on 13 taxa

The data sets obtained after the validation tests each contained 11 497 34 483 sequences depending on the species (Table 1). Mean sequence length was between 150 and 275 bp (Table 1), and maximum sequence length was between 477 and 997 bp (Table 1). The number of validated loci containing microsatellites ranged from 199 to 5791 (Table 2). We found that 3 28% of the validated loci corresponded to consensus sequences (Table 2), providing information about the polymorphism of sequences. Despite the heavy constraints imposed on primer design, primers were successfully designed for 94 1162 loci, even for species for which severe difficulties were previously encountered using traditional methods (Meglécz *et al.* 2004; Péténian *et al.* 2005; Dutech *et al.* 2007). The primers designed targeted various microsatellite motifs in each species, with the respective proportions of each motif differing among species (Table 2). Both the contrasted numbers and proportions of motifs found in the different organisms are not surprising with regard to the available literature (Lagercrantz *et al.* 1993; Toth *et al.* 2000; Meglécz *et al.* 2007; Richard *et al.* 2008).

Table 2 Description of the microsatellite libraries produced for the 14 tested species number of microsatellite loci validated by the QDD program; number of perfect (only one repeated motif) and compound (2 or more repeated motifs or interrupted microsatellites) microsatellite motifs among the loci for which PCR primers were designed by QDD; frequencies of the various types of perfect microsatellites Data are obtained from 454 runs in which 75 000 beads were loaded for each species, except *Apis mellifera* shotgun, enriched-125K, enriched-75K, enriched-50K and enriched-25K, for which 125 000, 125 000, 75 000, 50 000 and 25 000 beads were loaded, respectively

Name	Validated loci	% loci identified from several sequences	Design of primers		Motif type of perfect ms									
			Perfect ms	Compound	AC (%)	AG (%)	AAC (%)	AAG (%)	ACG (%)	AGG (%)	ACAT (%)	ATCT (%)	AT-based (%)	Others (%)
A mellifera (shotgun)	5157	3	1451	594	3	35	1	5	2	6	0	0	39	10
A mellifera (enriched-125K)	6230	22	1516	684	5	58	3	12	1	7	0	0	8	6
A mellifera (enriched-75K)	4674	21	1207	563	5	54	2	13	1	9	0	0	10	5
A mellifera (enriched-50K)	5025	21	1203	540	5	56	3	12	1	8	0	0	9	5
A mellifera (enriched-25K)	2519	18	656	290	5	55	3	12	1	8	0	0	11	5
Venturia canescens	2404	23	483	192	15	62	5	9	2	2	0	0	2	3
Euphydryas aurinia	1627	10	252	96	41	17	7	11	0	0	6	4	6	8
Pseudococcus viburni	1311	23	315	136	16	15	14	7	6	18	2	1	3	19
Diabrotica virgifera	1731	5	319	109	14	11	1	60	1	3	2	2	2	5
Bursaphelenchus xylophilus	199	28	71	23	20	45	11	13	0	3	0	0	0	8
Barbus meridionalis	2562	11	572	229	80	15	0	2	0	0	0	1	1	1
Danio rerio	5791	8	770	322	73	8	5	1	0	1	0	1	1	11
Gerbillus nigeriae	2048	3	163	78	36	19	10	1	0	3	3	9	2	17
Armillaria ostoyae	4015	6	235	69	20	10	22	10	4	6	0	1	6	20
Phytophthora alni	550	16	186	53	25	23	10	10	4	8	1	1	3	17
Festuca eskia	1474	8	475	151	31	35	8	10	1	6	0	1	4	4
Hirtella physophora	3552	24	809	353	16	24	9	18	2	11	1	2	6	10
Papaver rhoeas	1072	4	334	120	11	24	13	43	0	4	0	0	1	4

Discussion

Even with the stringent selection of loci imposed by the analysis parameters chosen, the numbers of microsatellite markers isolated with this new procedure were satisfactory for all the tested species. When compared to traditional isolation techniques, these numbers are much larger and were obtained with a much lower budget and within a shorter length of time. Indeed, whereas the isolation of microsatellites from an enriched library typically involves the screening and Sanger sequencing of a couple of hundred clones over a period of 5 weeks at a cost of more than US$ 5000, the process tested here allows the simultaneous screening of thousands of DNA fragments, over a 2-week period, at a cost of less than US$ 1500 (including expenses related to consumables and staff). The sequences produced are shorter than those obtained by Sanger sequencing, but this is not a problem for microsatellite marker development because the PCR products used for genotyping are usually 100 400 bp in size. Besides, this read-length constraint should be overcome with subsequent updates of 454 platforms. Comparing our results to the first studies using pyrosequencing to isolate microsatellites is not straightforward because (i) the way microsatellites were defined and the analysis methods used differ substantially, (ii) the organisms used to test the protocols are different and (iii) the total cost to generate a given amount of exploitable microsatellite sequences was generally not provided. However, the efficiency as evaluated from our results (between 1% and 8% of amplifiable markers in the obtained sequences; see methods for the definition of amplifiable markers) appears higher than in Abdelkrim *et al.* (2009) who reported around 0.1% of amplifiable microsatellites markers in their shotgun library and of the same order of magnitude as in Castoe *et al.* 2010 (~3.5% of amplifiable microsatellites in one species) and Santana *et al.* 2009 (between ~2% and ~5%).

Our tests revealed that the use of the multiplex DNA enrichment of our procedure appears highly advisable to optimize cost-efficiency of microsatellite isolation. Comparisons between enriched and shotgun *A. mellifera* libraries revealed two major advantages of enrichment. First, enrichment slightly increased the overall number of microsatellite loci isolated and reduced the proportion of unwanted motifs such as AT-based motifs (39 8%) that are likely to generate difficult amplification during genotyping. Counting the numbers of amplifiable optimal microsatellite markers (≥8 repetitions of perfect and not AT-rich motifs, i.e. the most valuable markers for the end-users) revealed that enrichment improved marker isolation efficiency by almost 300% (543 vs. 186 markers isolated in *A. mellifera* with and without enrichment) for an additional cost of around 10%. Moreover, the benefits of enrichment are likely much underestimated here because *A. mellifera* genome is rich in microsatellites (Solignac *et al.* 2007). Second, it increases the number of multiple reads obtained for a given microsatellite locus, which enables to design PCR primers targeting nonpolymorphic sequences flanking the microsatellite motif. Such an approach should decrease the probability of designing markers with a high percentage of null alleles because of mismatches between primers and polymorphic nucleotides in flanking regions that can occur in some individuals or populations. The analysis of the large data sets obtained, using programs like QDD, also enables to sort and discard loci that are likely to be found at multiple sites in the genome.

The validation of our method on 13 taxa did not reveal negative effects of enrichment on microsatellite isolation efficiency. Comparison of motif frequencies in the libraries generated here and in published genome sequences of closely related taxa (when available) shows that enrichment successfully favoured the most common motifs (Tables 2 and S1, Supporting information). Second, results do not indicate that lower isolation efficiency could be because of enrichment failure: we found no negative correlation between the percentage of unwanted motifs and the number of detected microsatellite loci ($r = 0.16$; $P > 0.05$). Lower yields in some species may rather be caused by technical issues (e.g. random manipulation effects or lower DNA quality of available samples) or lower abundances of microsatellites in genomes.

In conclusion, we demonstrated that our procedure coupling multiplex enrichment, pyrosequencing and sequence selection can be readily and successfully applied to a large variety of taxonomic groups (Table 2). The procedure is particularly cost-efficient as only a small part of a 454 GsFLX plate is needed to isolate high numbers of microsatellite markers. It is expected to speed up the acquisition of high-quality genetic markers for nonmodel organisms.

Acknowledgements

We thank M. Galan for useful comments on previous versions of the manuscript and S. Nielsen and J. Sappa for major improve ments to English grammar throughout the text. This work was supported by the AIP BioRessources 'EcoMicro' grant from the French *Institut National de la Recherche Agronomique* (INRA), permanent institutional support from Montpellier SupAgro, University Aix Marseille I, INRA and the R&D budget of Genoscreen (Lille, France).

References

Abdelkrim J, Robertson BC, Stanton JAL, Gemmell NJ (2009) Fast, cost-effective development of species-specific microsatellite markers by genomic sequencing. *BioTechniques*, **46**, 185 191.

Allentoft ME, Schuster SC, Holdaway RN *et al.* (2009) Identification of microsatellites from an extinct moa species using high-throughput (454) sequence data. *BioTechniques*, **46**, 195 200.

Castoe TA, Poole AW, Gu W *et al.* (2010) Rapid identification of thousands of copperhead snake (*Agkistrodon contortrix*) microsatellite loci from modest amounts of 454 shotgun genome sequence. *Molecular Ecology Resources*, **10**, 341 347.

Dutech C, Enjalbert J, Fournier E *et al.* (2007) Challenges of microsatellite isolation in fungi. *Fungal Genetics and Biology*, **44**, 933 949.

Kijas JMH, Fowler JCS, Garbett CA, Thomas MR (1994) Enrichment of microsatellites from the *Citrus* genome using biotinylated oligonucleotide sequences bound to streptavidin-coated magnetic particles. *BioTechniques*, **16**, 656 662.

Lagercrantz U, Ellegren H, Andersson L (1993) The abundance of various polymorphic microsatellite motifs differs between plants and vertebrates. *Nucleic Acids Research*, **21**, 1111 1115.

Meglécz E, Petenian F, Danchin E *et al.* (2004) High similarity between flanking regions of different microsatellites detected within each of two species of Lepidoptera: *Parnassius apollo* and *Euphydryas aurinia*. *Molecular Ecology*, **13**, 1693 1700.

Meglécz E, Anderson SJ, Bourguet D *et al.* (2007) Microsatellite flanking region similarities among different loci within insect species. *Insect Molecular Biology*, **16**, 175 185.

Meglécz E, Costedoat C, Dubut V *et al.* (2010) QDD: a user-friendly program to select microsatellite markers and design primers from large sequencing projects. *Bioinformatics*, **26**, 403 404.

Péténian F, Meglécz E, Genson G, Rasplus JY, Faure E (2005) Isolation and characterization of polymorphic microsatellites in *Parnassius apollo* and *Euphydryas aurinia* (Lepidoptera). *Molecular Ecology Notes*, **5**, 243 245.

Richard GF, Kerrest A, Dujon B (2008) Comparative genomics and molecular dynamics of DNA repeats in Eukaryotes. *Microbiology and Molecular Biology Reviews*, **72**, 686 727.

Rozen S, Skaletsky H (2000) Primer3 on the WWW for general users and for biologist programmers. *Methods in Molecular Biology*, **132**, 365 386.

Santana QC, Coetzee MPA, Steenkamp ET *et al.* (2009) Microsatellite discovery by deep sequencing of enriched genomic libraries. *BioTechniques*, **46**, 217 223.

Solignac M, Mougel F, Vautrin D, Monnerot M, Cornuet J-M (2007) A third-generation microsatellite-based linkage map of the honey bee, *Apis mellifera*, and its comparison with the sequence-based physical map. *Genome Biology*, **8**, R66.

Sunnucks P (2000) Efficient genetic markers for population biology. *Trends in Ecology and Evolution*, **15**, 199 203.

Toth G, Gaspari Z, Jurka J (2000) Microsatellites in different eukaryotic genomes: survey and analysis. *Genome Research*, **10**, 967 981.

Supporting Information

Additional supporting information may be found in the online version of this article.

Table S1 Proportions of microsatellite motifs in selected genomes.

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting information supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.