Published online 20 February 2007

Nucleic Acids Research, 2007, Vol. 35, No. 6 ***e44***
doi:10.1093/nar/gkm070

SNP discovery by mismatch-targeting of Mu transposition

Luisa Orsini[1], Maria Pajunen[2], Ilkka Hanski[1] and Harri Savilahti[2,3,*]

[1]Metapopulation Research Group, Department of Biological and Environmental Sciences, PO Box 65, and [2]Research Program in Cellular Biotechnology, Institute of Biotechnology, PO Box 56, FIN-00014, University of Helsinki, Finland and [3]Division of Genetics and Physiology, Department of Biology, FIN-20014, University of Turku, Finland

Received September 15, 2006; Revised January 18, 2007; Accepted January 23, 2007

ABSTRACT

Single nucleotide polymorphisms (SNPs) represent a valuable resource for the mapping of human disease genes and induced mutations in model organisms. SNPs may become the markers of choice also for population ecology and evolutionary studies, but their isolation for non-model organisms with unsequenced genomes is often difficult. Here, we describe a rapid and cost-effective strategy to isolate SNPs that exploits the property of the bacteriophage Mu transposition machinery to target mismatched DNA sites and thereby to effectively detect polymorphic loci. To demonstrate the methodology, we isolated 164 SNPs from the unsequenced genome of the Glanville fritillary butterfly (*Melitaea cinxia*), a much-studied species in population biology, and we validated 24 of them. The strategy involves standard molecular biology techniques as well as undemanding MuA transposase-catalyzed *in vitro* transposition reactions, and it is applicable to any organism.

INTRODUCTION

Single nucleotide polymorphisms (SNPs) represent the most widespread type of sequence variation in genomes (1) and provide the most commonly used genetic markers for the mapping of human disease genes (2) as well as experimentally induced mutations in model organisms (3). Disciplines such as population ecology and conservation and evolutionary genetics would equally benefit from SNPs as genetic markers, but their use for such purposes has been limited due to the expenses and technical difficulties involved in the currently available SNP isolation strategies for non-model organisms. Any methodology that would streamline the SNP discovery process, particularly for non-model organisms, would be highly desirable (4,5).

Typical direct SNP discovery strategies (6,7) involve sequencing of locus-specific amplification (LSA) products from multiple individuals or sequence determination of expressed sequence tags (EST-sequencing). Other direct strategies include whole genome (WGSS) and reduced representation (RRSS) shotgun sequencing approaches. If comparative sequence data are available in public or other databases, various sequence comparison algorithms that identify nucleotide differences provide an alternative means to empirically discover SNPs (8).

Indirect SNP discovery strategies include a prescreening phase prior to sequence determination, and these methodologies detect heteroduplexes on the basis of mismatch-induced altered DNA characteristics. Physical differences are exploited in electrophoretic analyses such as single-strand conformational polymorphism (SSCP) that relies on conformation-dependent allele-specific migration differences of single strands (9). Similarly, altered melting behavior of mismatch-containing DNA fragment can be utilized to detect SNPs (10–12). Chemical differences can also be utilized for SNP detection. In principle, any reagent that specifically recognizes and cleaves mismatched DNA can be used for the detection (13), and heteroduplex-cleaving chemicals (14) or proteins (15) have been used for the purpose. Recently, a novel DNA-cleaving reagent became available when it was shown that Mu transposition preferentially targets mismatched sites in DNA (16). This proof of principle study established the mismatch-targeting methodology and indicated, using a known polymorphic test fragment, that mutations indeed can be detected by the use of transposon approach (16).

The present study has been stimulated by research on the Glanville fritillary butterfly (*Melitaea cinxia*). This species and its large metapopulation in Finland have become a much-studied model system in population biology (17). Adding a strong genetic component into the existing ecological context would be highly desirable, but the paucity of suitable genetic markers has hampered the progress towards this goal. In particular, the development

*To whom correspondence should be addressed. Tel: +358 9 191 59516; Fax: +358 9 191 59366; Email: harri.savilahti@helsinki.fi

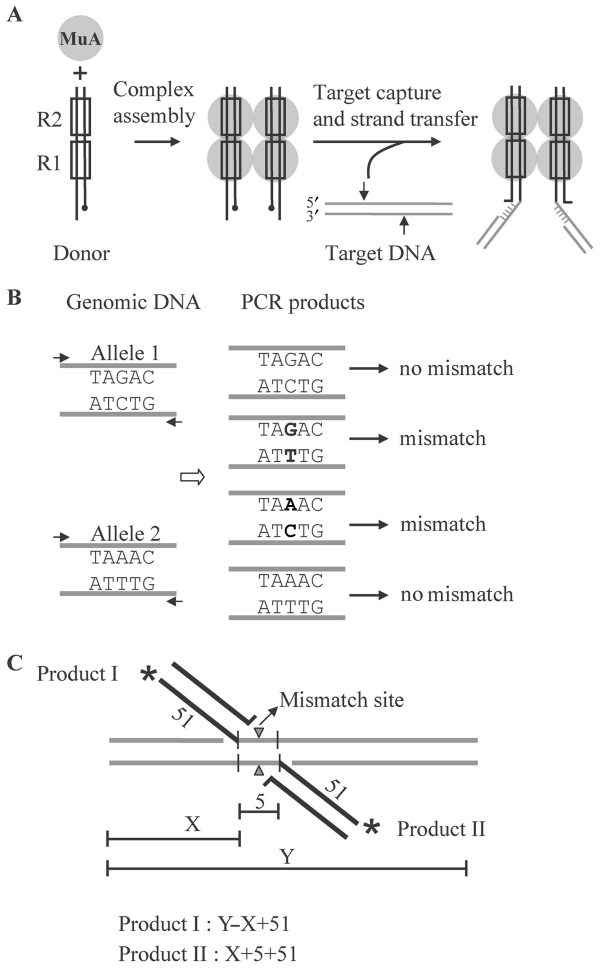

Figure 1. Mismatch targeting of Mu transposition. (**A**) Outline of the Mu transpositional recombination steps used in this study. MuA transposase protein assembles two transposon end segments into a tetrameric DNA transposition complex. This complex captures the target DNA and executes the strand transfer reaction, during which the transposon DNA is joined into the target in a concerted reaction involving a 5-bp stagger, and the target DNA strands are simultaneously cleaved. R1 and R2 (rectangles) denote MuA transposase-binding sites. The arrows indicate the 5-bp staggered locations for strand transfer on the two strands. When mismatched sites are present in the target DNA, nearly 90% of the strand transfers occur at these sites (16). (**B**) If genomic DNA contains at least two alleles within a specified DNA region, amplification of that region by PCR produces DNA duplexes that contain mismatches. Such a situation arises when the region is amplified from a heterozygous individual or from a sample that combines DNA from two or more individuals representing different allelic variants. In this example, mismatched nucleotides are shown in bold. (**C**) Lengths of the DNA strands within the transposition product. Transposon DNA is shown in black and target DNA in gray. Numbers indicate known lengths (in nucleotides), and labeled reaction products are indicated with asterisks. Two formulas for the calculation of the product lengths are shown at the bottom.

of effective microsatellite markers for Lepidoptera species, including *M. cinxia*, has turned out to be difficult (18). Possible reasons for this may involve a high degree of variation close to the microsatellite loci as well as the presence of duplicated genomic regions or several copies of mobile elements (19). Considering the above difficulties, other types of genetic markers are needed, and for many purposes SNPs represent an attractive alternative.

Here, we adopted the methodology of Mu transposition to detect mismatches in DNA (16) and developed a strategy to isolate SNP markers from uncharted genomes. The methodology exploits the bacteriophage Mu DNA transposition machinery, the critical components of which include a tetramer of MuA transposase and two transposon end segments (20,21). The assembly of this machinery and subsequent transposase-catalyzed reaction steps (Figure 1A) can be reconstituted in a simple *in vitro* reaction that includes transposon DNA (a short Mu genome right-end segment suffices), MuA transposase and target DNA as the only macromolecular components (20). This minimal *in vitro* reaction has recently been used in a number of advanced molecular biology, protein engineering and genomics applications (22–27), and it has become evident that many other novel applications can be tackled with this technology.

MATERIALS AND METHODS

DNA techniques

Adult Glanville fritillary butterflies were collected from several locations on the Åland Islands (Finland), and their genomic DNA was isolated using the Nucleo spin tissue extraction kit (Mackerey-Nagel). Plasmids were propagated in *E. coli* DH5α (Invitrogen) and isolated using appropriate Qiagen kits. Standard DNA techniques were performed as described (28). MuA transposase (MuA) was purified in collaboration with Finnzymes (Espoo, Finland) as described (23,29). Origin of other proteins, oligonucleotides and reagents are listed in Table S1. The MM1141 oligonucleotide was radiolabeled at the 5′-end with [γ-^{33}P]ATP using T4 polynucleotide kinase (28). The ^{33}P-labeled MM1141 was purified and annealed with unlabeled MM1138 to generate a radiolabeled Mu end DNA fragment as described (20). DNA-modifying enzymes were used as recommended by the suppliers. A MegaBace 1000 sequencer (GE Healthcare) with Big Dye Terminator chemistry was used for sequencing. Sequencing ladders were produced using the Sequenase 2.0 sequencing kit (USB).

Melitaea cinxia genomic DNA library

DNA from four individuals (#1, #2, #3 and #4) was pooled, digested partially with Sau3AI and size-fractionated on a preparative 1.5% SeaPlaque (Cambrex) agarose gel. Electroeluted (28) fragments (150–1000 bp) were cloned into the BamHI site of plasmid pUC19 (New England Biolabs) to yield a library of ~20 000 clones. We decided to use four individual butterflies on the basis of the prediction that 1000 bp of genomic DNA would contain one variable nucleotide between two individuals on average (30). With four individuals, the probability of detecting variation within 250-bp fragments should then be high. While this estimate was not based on a known variation, it proved to be appropriate in practice.

Sequence comparison and primer design

Inserts were sequenced on both strands with appropriate primers (GeneBank AN: DQ389519–DQ389533), and the data were used for blast searches (31) in the NCBI and Silkworm Knowledgebase (Silk DB-http://silkworm.genomics.org.cn/index.jsp) public databases. The PRIMER3 program (32) was used to design primer pairs for genomic DNA amplification (targeted product size range 150–1000 bp).

Production of genomic DNA fragments

DNA fragments from the Glanville fritillary genome were produced by PCR amplification using a MBS 0.2G thermal cycler (Hybaid). The optimal annealing temperature was defined experimentally for each primer pair using the temperature gradient option of the cycler. Each amplification reaction (20 µl) contained 20–30 ng genomic DNA, 0.5 µM each primer, 200 µM each dNTPs and, 2.5 mM $MgCl_2$, 20 ng BSA and 0.1 U Taq DNA polymerase. An initial denaturation step (5 min at 95°C) was followed by 35 cycles of amplification with 1 min at 94°C, 1 min at the annealing temperature and 1.5 min at 72°C. In these reaction conditions, depending on the length of the PCR product, either nucleotides or primers become a limiting factor. Thus, heteroduplexed DNA fragments will be formed (in the presence of allelic variation), and no extra annealing step is required. Each PCR product was purified from several parallel reactions using a Gen-Pak FAX (Waters) anion exchange column. Each product was ethanol precipitated and resuspended in TEN buffer (10 mM Tris-HCl, pH 7.5, 0.5 mM EDTA, 50 mM NaCl). Alternatively, PCR fragments were purified using GFX purifying kit (GE Healthcare). The human control fragment was amplified from the HLA region of an anonymous individual as described (16).

Production of mismatch-containing model target fragments

Initially, three variants of a known 1994-bp DNA fragment (MuA gene cloned in a plasmid vector) were amplified separately. These variants represent wild-type DNA and two different point mutations, G785A and T1102C. Each amplification reaction (50 µl) contained 20 ng plasmid DNA template, 0.5 µM each primer, 200 µM each dNTPs and 1 U Phusion DNA polymerase (in Phusion HF buffer). An initial denaturation step (2 min at 98°C) was followed by 35 cycles of amplification with 0.5 min at 98°C, 1 min at 57°C and 1 min at 72°C. Each PCR product was purified using QIAquick PCR purification kit (Qiagen). Each mismatch-containing target duplex was generated by initially mixing the wild-type and mutant fragment (in TEN buffer, molar ratio 1:1). Denaturation (2 min at 95°C) was followed by a slow cooling to room temperature. A size marker for the polyacrylamide gel electrophoresis (Figure S1A) was generated by radiolabeling NdeI-digested plasmid pLEB620 (33) at the 5′-ends with [γ-^{33}P]ATP using T4 polynucleotide kinase (28) and further digesting the plasmid with XhoI.

Mismatch-targeting analysis by *in vitro* transposition reaction

Initially, the reaction was incubated for 1 h at 30°C in the absence of target DNA and divalent metal ions to allow the assembly of the transposition machinery. Next, target DNA (3.5 µl in TEN buffer) was added to the reaction mixture (20.5 µl) and the incubation was continued for 10 min to allow target capture. To activate catalysis, $MgCl_2$ (1 µl of 250 mM stock) was subsequently added, and the incubation was continued for further 10 min. At this final incubation stage, the reaction (25 µl) contained 50 nM ^{33}P-labeled Mu end DNA fragment, 260 ng target DNA, 116 nM (0.2 µg) MuA, 25 mM Tris-HCl, pH 8.0, 100 µg/ml BSA, 15% (w/v) glycerol, 15% (v/v) DMSO, 0.05% (w/v) Triton X-100, 119 mM NaCl and 10 mM $MgCl_2$. The reactions were terminated by freezing in liquid nitrogen, and reaction products were analyzed by denaturing 7 M urea, 6% polyacrylamide gel electrophoresis and autoradiography as described. (34). Intensities of the transposition reaction products were quantified (Figure S1B) with the Aida 3.44 program (Raytest).

Determination of sequence variation among individuals

The 16 genomic target fragments that yielded an evident banding pattern in the mismatch-targeting analysis were processed for method validation by sequencing both DNA strands with appropriate primers as follows. Each fragment was produced separately from the DNA of four individuals (#1, #2, #3, #4) and cloned into plasmid pGEM-T easy (Promega) for sequence determination. From four to six independent clones per each individual were sequenced to reach a high probability of gaining information from both sister chromosomes (GeneBank AN: DQ389251–DQ389518, DQ389534–DQ389576). Subsequently, the data were compiled and used to design 24 SNP probes for genotyping.

Genotyping

Eight butterflies, each collected from a different population on the Åland Islands as well as the four individuals used in the original genomic library screen were genotyped for the 24 newly identified SNPs. The SNP genotyping was performed using the SnuPe kit on a MegaBace 1000 sequencer, following the instructions of the supplier (GE Healthcare).

RESULTS

Genomic amplification

We made a genomic library of the Glanville fritillary and screened 134 clones for inserts by restriction analysis, of which 102 (size range 150–1000 bp, Table S2) were sequenced. The data obtained were used in Blast searches against the *Bombyx mori* genome and against the arthropod sequences available in the NCBI public database to identify similarities. All the 102 sequences yielded significant matches with *B. mori*. In addition, half of them yielded significant similarity with other insects,

demonstrating the authenticity of the cloned butterfly DNA (a representative sample in Table S3). A total of 122 primer pairs were designed and used for genomic amplification (1–4 per locus), and a DNA sample pooled from four individual butterflies was used as template. Forty-four pairs (36%) generated amplification products that appeared as a single, predictable-size band in an agarose gel analysis. The remaining primer pairs yielded no detectable products (19 pairs) or amplified more than one product (59 pairs). Several parallel amplification reactions were next used to produce 32 genomic fragments, and these fragments were purified by anion exchange chromatography.

Detection of variation by *in vitro* transposition

To identify SNPs detectable as mismatched DNA sites, we performed 32 Mu *in vitro* transposition reactions (Figure 1A). Each reaction included radioactively labeled Mu R-end segment as a donor and one of the 32 purified PCR fragments (Figure 1B) as a target (pooled DNA sample from four individuals). Reaction products (Figure 1C) were analyzed on a denaturing polyacrylamide gel and visualized by autoradiography (Figure 2). Sixteen reactions yielded a pattern of relatively evenly distributed bands, evidently representing integrations at various positions along the target DNA, a typical reaction profile in cases where no mismatches are present in the target DNA (22,34). We conclude that these 16 fragments probably do not contain allelic variation, and they were not studied further. In contrast, the remaining 16 reactions generated easily recognizable discrete bands on top of a more evenly distributed background band patterning, indicating presence of mismatched DNA sites and revealing variation. A mismatch-containing control fragment (16), amplified from the highly polymorphic human HLA locus, produced a comparable band pattern (Figure 2, lane HLA).

Verification of variation by sequencing

The 16 variation-indicating genomic segments were examined for allelic differences at the sequence level. All the segments were cloned individually from each of the four butterflies used for the genomic library construction, and up to six independent clones were sequenced to reach a reasonable probability of obtaining data from both sister chromosomes. Three fragments were evidently doublets, representing two independent loci, and they were not studied further. The remaining 13 fragments each yielded information from a single locus (Table S4).

Mismatch analysis versus observed variation

The correspondence between the autoradiographic data and sequence variation was next examined in detail for several fragments. Below, we present a critical evaluation of three representative fragments to highlight the performance of the methodology under different levels of heterozygosity and variation. For each fragment, DNA was analyzed both from each of the four individuals separately and from a pooled sample of these four individuals (Figure 3).

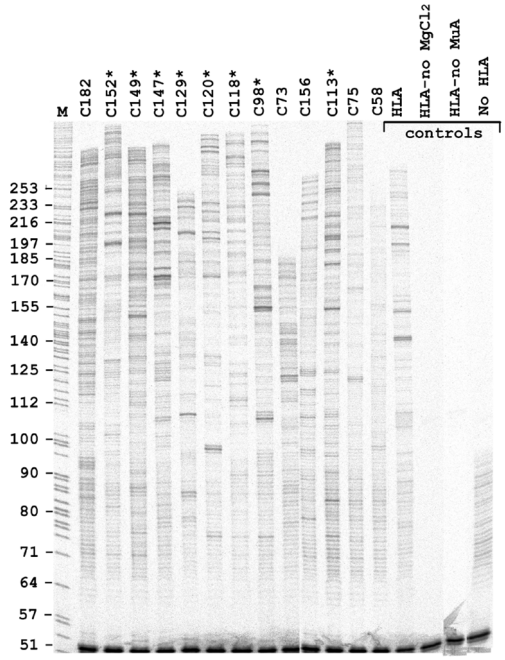

Figure 2. Mismatch-targeting analysis. Strand transfer products of *in vitro* transposition reactions were analyzed by denaturing polyacrylamide gel electrophoresis and autoradiography. Each lane represents a unique target DNA fragment amplified using as a template a genomic DNA sample pooled from four butterflies. Asterisks highlight distinctive band patterns and indicate clones selected for further analyses. M: molecular marker; HLA: positive control fragment (from human HLA locus); HLA-no $MgCl_2$: magnesium chloride omitted; HLA-no MuA: MuA transposase omitted; No HLA: transposition reaction without the addition of extra target DNA (in this case the donor DNA fragment serves as a target).

Figure 3A illustrates a case where all the individuals apparently are homozygous with respect to the variable nucleotides. Three individuals (#2, #3 and #4) evidently belong to the same haplotype. The fourth individual (#1) represents a second haplotype with a single nucleotide substitution in ten positions and a 3-nt insertion in one position. For the three individuals, an identical and relatively even, homozygosity-indicating band pattern was expected and observed. The fourth individual generated a similar even pattern with a slight overall shift in the middle, evidently reflecting the 3-nt insertion. When the pooled DNA sample was analyzed, a distinctive and strikingly symmetrical band pattern emerged (see Table S5 for the calculation of fragment sizes).

Figure 3B shows an example in which the banding pattern varies among the four individuals, reflecting allelic variation between the sister chromosomes of heterozygous individuals. Within this genomic segment, all the detectable variation is revealed by the analysis of the four individuals separately, illustrating a situation where no extra information is gained by the use of a pooled DNA sample.

Figure 3C shows a very complex situation, where a high degree of variation generates an elaborate pattern of bands. While this complexity, particularly the indels,

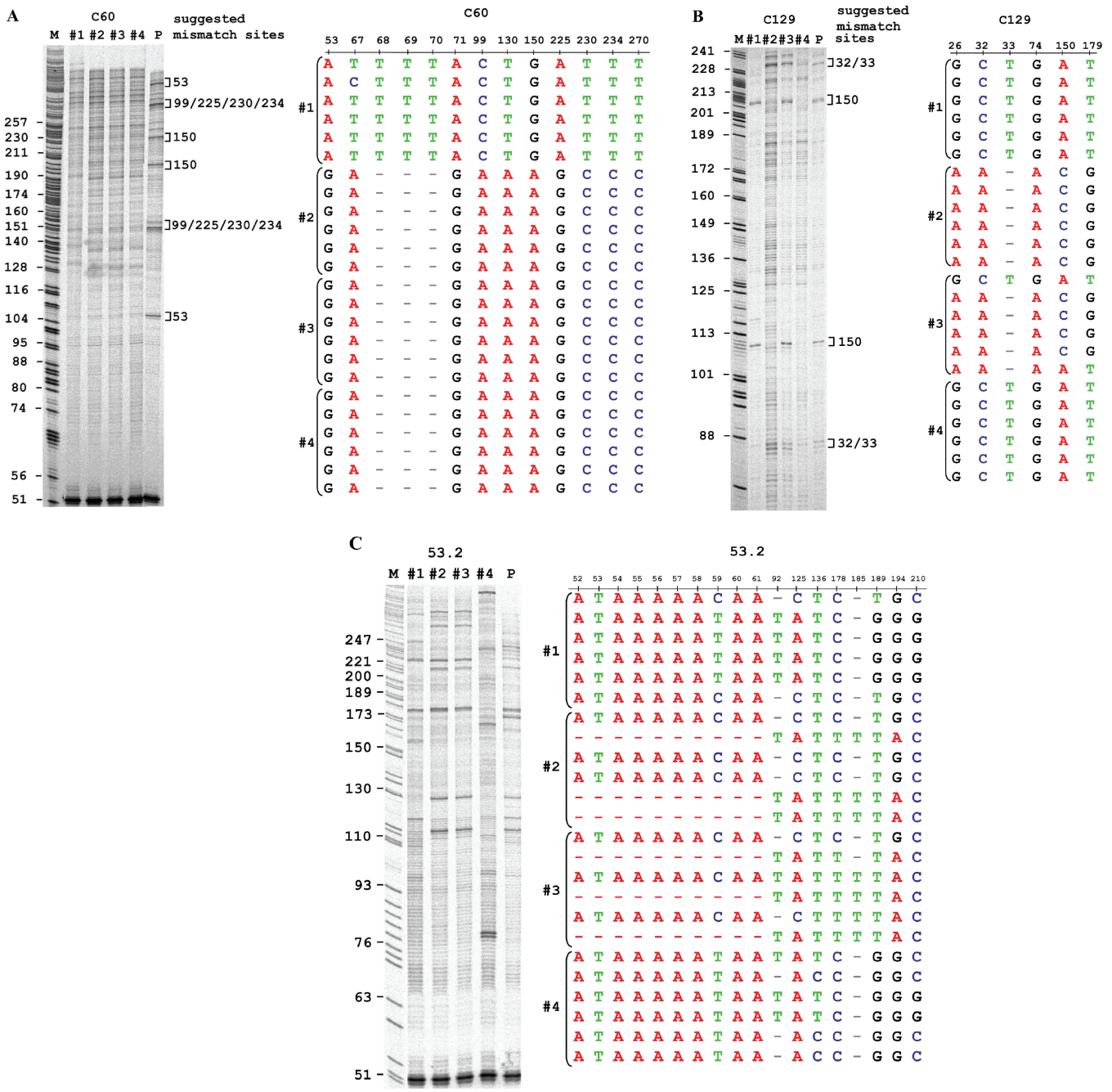

Figure 3. Correspondence between the autoradiographic banding pattern and the observed sequence variation. M: marker; #1, #2, #3 and #4: four individual DNA samples; P: a pooled DNA sample of the four individuals. (**A**) Individuals are apparently homozygous with regard to the variable nucleotides. The observed banding pattern results from transposon integrations at or close to any of the mismatched sites at nucleotide positions 53, 99, 150, 225, 230, 234 and 270 (suggested mismatch sites, Table S4). (**B**) Individuals show different levels of heterozygosity versus homozygosity. (**C**) An example of a high degree of variation including indels. The autoradiography versus sequence data sets are mutually consistent. When interpreting the data, note that some of those variant nucleotides that appear only once in the sequence compilation probably do not reflect actual genomic variation but represent either sequencing errors or PCR-cloning-generated spurious mutations. Also, the sequence analysis may not have detected all variant nucleotides in certain heterozygous individuals for statistical reasons, as the probability of not obtaining data from both sister chromosomes is 2% with six independently examined clones.

challenges accurate nucleotide-level interpretation of the data, the banding pattern very clearly indicates similar or possibly identical genotypes in heterozygous individuals (e.g. individuals #2 and #3), and in general the autoradiograph portrays a picture of extensive variation.

SNPs and genotyping

We identified 13 variable loci that were represented by DNA fragments with an average length of 287 bp (184–331 bp), totaling 3738 bp of genomic sequence. Within these regions we discovered 164 SNPs

Table 1. Polymorphisms detected within the genome of the Glanville fritillary butterfly

Genomic region	Nucleotide substitutions	Indels		Total number of SNPs	Probe design
		1 bp	>1 bp		
30.1	9	0	0	9	3
42.1	3	0	1	4	2
53.2	8	2	1	11	3
3N	6	1	2	9	2
30N	1	0	0	1	1
C60	10	0	1	11	2
C98	9	0	1	10	4
C113	17	1	1	19	2
C120	10	0	1	11	2
C129	6	0	0	6	2
C152	1	0	0	1	1
C118	46	2	4	52	0
C147	17	0	3	20	0
Total	143	6	15	164	24

The total number of nucleotide substitutions and indels are listed for 13 unique genomic regions. The total number of the SNP probes designed is also shown.

(Table 1 and Table S6), indicating high intrinsic power of the methodology for the selection of polymorphisms. Twenty-four of the SNPs were suitable for probe design and subsequent genotyping (Table S4). These SNPs were distributed along 11 unique genomic segments with a total length of 3140 bp. Variation within this pre-selected DNA sample was remarkable, 29 SNPs per 1000 bp. To validate the proper functioning of the probe primers, 12 butterflies were collected from 12 randomly selected populations on the Åland Islands and genotyped for the 24 SNPs. In each case, we detected two alleles that were consistent with those identified by sequence analysis.

Known mismatches in longer fragments

Because of the high level of variation, mismatch analysis of relatively short fragments appeared to be optimally suited for the Glanville fritillary. However, in species where much less variation is present, it would be desirable to analyze larger segments of DNA. To test whether the Mu mismatch-targeting strategy would be able to detect a single point mutation in a longer DNA fragment, we generated two 1994-bp model fragments, each containing a mismatched nucleotide pair in a known position (see Methods). These two fragments were then used as targets in the mismatch analysis (Figure S1). With both of these mismatch-containing fragments, the autoradiograph and density scan of the signals revealed an expected band pattern representing two correct size fragments. These data indicate that the methodology is able to detect single mismatches even if they are present in a longer fragment (at least up to 2 kb).

DISCUSSION

Single nucleotide polymorphisms provide the markers of choice for evolutionary, ecological and conservation studies (5). The ease with which SNP data can be modeled as well as the abundance of SNPs in genomes make them ideal for the study of population histories (4). A major limiting factor for their use for non-model organisms in population biology has been the lack of an efficient and cost-effective method to isolate new markers. The mismatch-targeting of Mu transposition-based strategy described in this article has the potential to solve this problem. Important for many researchers in population genetics and evolutionary biology, this method requires no special facilities over standard molecular biology laboratory.

The present methodology involves undemanding cloning and sequencing steps, yielding data for the design of genome-specific primers. In this study, a third of the designed primer pairs amplified a single PCR product, and the rest of them failed in amplification or amplified several products. While some of the amplification problems may have been caused by sub-optimally designed primer pairs, we suspect that some of these failures may reflect substantial variation among individuals and/or stretches of sequence similarity in different loci. In general, variation within primer-binding sites may influence the amplification, and large indels are expected to generate several fragments. In addition, duplicated or otherwise similar but not identical genome regions as well as multiple copies of mobile elements can generate a complex set of amplification products. The fact that a high percentage of primer pairs generated multiple PCR products or failed to generate products in our study may relate to the exceptional difficulties encountered in the development of microsatellite markers for Lepidoptera species, including the Glanville fritillary. Indeed, a high level of variation within the flanking regions of microsatellites has hampered their use as markers (35,36).

Fragment length appears not to be very critical for the present methodology, as in a preliminary phase of this study, DNA fragments up to 1.3 kb in size were successfully analyzed for the presence of variation in the Glanville fritillary (data not shown). Thereafter, most of the analyzed butterfly DNA fragments were targeted to fall within the 250–350-bp size range for convenience: such fragments are short enough for straightforward genomic amplification, sequencing can be accomplished with one primer and optimal separation of transposition reaction products is achieved. Nevertheless, as shown with model DNA fragments (Figure S1), a single mismatched nucleotide pair can readily be detected even when it is present in a 2-kb fragment.

Mu mismatch-targeting can be easily visualized by the use of electrophoresis and autoradiography. The two transposon ends integrate simultaneously into each of the target DNA strands (Figure 1A), generating two complementary products (Figure 1C). Hence, the symmetrical banding pattern in autoradiographs serves as a built-in quality measure, discriminating against any potential artifacts. The lengths of the transposition reaction products can be estimated with a reasonable accuracy by the use of molecular size markers, although a degree of sequence-specific variation in migration does exist among single-strands. In most cases, the targeted mismatch is located in the middle of the 5-bp target region core (16), generating easily interpretable banding patterns

(Figure 1C, Table S5), although some targeting into nearby nucleotides may also occur (16).

We found that Mu-mediated integration can detect many, but not necessarily all, SNPs present within a particular DNA fragment. Thus, the autoradiographic data will underestimate the actual variation in cases where many variable nucleotides are present within a single genomic fragment. As the exact targeting mechanism of Mu transposition is currently not known, it is unclear why some sites are less effective than others, and what might be the maximum number of simultaneously identifiable mismatched sites within a given fragment. A suggestion that the machinery samples a large number of potential target sites before integration (16) is consistent with our data, but the mechanism of the site-discrimination process remains to be elucidated. Remarkably, the Mu machinery can mediate transposition at detectable levels into a mismatched site in the presence of 300 000-fold excess of non-mismatch sites, and all single nucleotide mismatch types as well as longer mismatches (at least up to 5 nt) target efficiently (16). In summary, the currently available data (16, this study) suggest that Mu transposition never fails to detect a single mismatch within a fragment, and some mismatches may become non-detectable only in fragments where they are present in a combination with those that can be detected, generating favorable circumstances for SNP discovery.

The Mu-mediated SNP discovery process discriminates effectively against invariant regions and detects variation-containing fragments with 100% efficiency. Therefore, sequencing can be focused on those regions where one will surely find polymorphic sites, thus avoiding massive and expensive sequencing efforts. On the other hand, too much variation is often problematic for primer probe design, and such problems can be avoided by choosing for sequencing only those fragments that show relatively few bands in the autoradiograph. Here, we selected 13 fragments with different degrees of variation, two of which were too variable for primer probe design.

Another advantage of the present methodology is the possibility to label the transposon DNA, alleviating the need to label each target fragment separately. Although we used radioactive labeling, non-radioactive protocols could be applicable as well. The benefit is that the labeled transposon reagent could be stored for extended periods of time for future use. The lack of apparent fragment size upper limit as such and the possibility to locate mutations with certain accuracy are clear advantages over methods that rely on conformational differences (SSCP and DGGE, see ref. 37 and references therein). However, with longer DNA fragments, gel resolution becomes a more pronounced issue, but similar problems apply to all methodologies that require resolution of different length DNA molecules. Also, the presence of indels may complicate the analysis, but this is a common problem among almost all currently available methods, excluding certain direct DNA sequencing approaches (6,7).

Of the currently available techniques, those that rely on enzymatic DNA cleavage agents, such as CEL I (38,39), are most closely related to the described Mu strategy. Yet, certain key differences exist. (i) In comparison to CEL I, Mu methodology does not require labeling of the target DNA; therefore, the labeling costs are minimized. (ii) The reaction products of CEL I cleavage are shorter than the labeled (target) DNA substrate. In contrast, the Mu methodology generates labeled products that are longer than the labeled (donor) DNA substrate, yielding favorable circumstances with regard to the signal to noise ratio. (iii) The detection of mutations very close to the ends of fragments is difficult with enzymatic mutation detection technologies, including CEL I. Because the Mu transposition product contains 51 extra nucleotides derived from the donor DNA (Figure 1), mutations located close to the end of the fragment are detectable by standard gel assays.

The methodology we describe here functions robustly, but some improvements may be envisioned. For example, the initial cloning step may not be necessary, as arbitrary priming and linker ligation-mediated protocols for genomic amplification are available (40,41). We purified the genomic PCR fragments by the use of chromatography, but any PCR purification method should be applicable. In fact, we tested one commercial kit (see Methods) for the purpose and the results compared favorably with those obtained with chromatographically purified fragments. In addition, many types of advanced technologies, including capillary electrophoresis and automation to generate a high-throughput environment, could be linked with the present methodology. Considering the numerous advantages, the mismatch-targeting of Mu transposition-based strategy described in this paper has the potential to become the favored approach to develop SNP markers for non-model organisms.

ACKNOWLEDGEMENTS

We thank Toshka Nyman and Pirjo Rahkola for excellent technical assistance and Tiina Luukkainen for the longer mismatch-containing targets. This project has been funded by the Academy of Finland Finnish Centre of Excellence Program 2000–2005 (grant to IH) and the Finnish National Technology Agency (TEKES) Neo-Bio Program 2001–2005 (grant to H.S.). Funding to pay the Open Access publication charge was provided by TEKES.

Conflict of interest statement. None declared.

REFERENCES

1. Collins,F.S., Brooks,L.D. and Chakravarti,A. (1998) A DNA polymorphism discovery resource for research on human genetic variation. *Genome Res.*, **8**, 1229–1231.
2. Consortium,T.I.H. (2005) A haplotype map of the human genome. *Nature*, **437**, 1299–1320.
3. Berger,J., Suzuki,T., Senti,K.A., Stubbs,J., Schaffner,G. and Dickson,B.J. (2001) Genetic mapping with SNP markers in Drosophila. *Nat. Genet.*, **29**, 475–481.
4. Brumfield,R.T., Beerli,P., Nickerson,D.A. and Edwards,S.V. (2003) The utility of single nucleotide polymorphisms in inferences of population history. *Trends Ecol. Evol.*, **18**, 249–256.
5. Seddon,J.M., Parker,H.G., Ostrander,E.A. and Ellegren,H. (2005) SNPs in ecological and conservation studies: a test in the Scandinavian wolf population. *Mol. Ecol.*, **14**, 503–511.

6. Suh,Y. and Vijg,J. (2005) SNP discovery in associating genetic variation with human disease phenotypes. *Mutat Res.*, **573**, 41–53.
7. Twyman,R.M. (2004) SNP discovery and typing technologies for pharmacogenomics. *Curr. Top. Med. Chem.*, **4**, 1423–1431.
8. Guryev,V., Berezikov,E. and Cuppen,E. (2005) CASCAD: a database of annotated candidate single nucleotide polymorphisms associated with expressed sequences. *BMC Genomics*, **6**, 10.
9. Orita,M., Iwahana,H., Kanazawa,H., Hayashi,K. and Sekiya,T. (1989) Detection of polymorphisms of human DNA by gel electrophoresis as single-strand conformation polymorphisms. *Proc. Natl. Acad. Sci. U.S.A.*, **86**, 2766–2770.
10. Fisher,S. and Lerman,L.S. (1979) Length-independent separation of DNA restriction fragments in two-dimensional gel electrophoresis. *Cell*, **16**, 191–200.
11. Vijg,J. and van Orsouw,N.J. (1999) Two-dimensional gene scanning: exploring human genetic variability. *Electrophoresis*, **20**, 1239–1249.
12. Xiao,W. and Oefner,P.J. (2001) Denaturing high-performance liquid chromatography: a review. *Hum. Mutat.*, **17**, 439–474.
13. Goldrick,M.M. (2001) RNase cleavage-based methods for mutation/SNP detection, past and present. *Hum. Mutat.*, **18**, 190–204.
14. Ellis,T.P., Humphrey,K.E., Smith,M.J. and Cotton,R.G.H. (1998) Chemical cleavage of mismatch: a new look at an established method. *Hum. Mutat.*, **11**, 345–353.
15. Till,B.J., Burtner,C., Comai,L. and Henikoff,S. (2004) Mismatch cleavage by single-strand specific nucleases. *Nucleic Acids Res.*, **32**, 2632–2641.
16. Yanagihara,K. and Mizuuchi,K. (2002) Mismatch-targeted transposition of Mu: a new strategy to map genetic polymorphism. *Proc. Natl. Acad. Sci. U.S.A.*, **99**, 11317–11321.
17. Ehrlich,P.R. and Hanski,I. (2004) *On the Wings of the Checkerspots: A Model System for Population Biology*. Oxford University Press, New York.
18. Neve,G. and Meglecz,E. (2000) Microsatellite frequencies in different taxa. *Trends Ecol. Evol.*, **15**, 376–377.
19. Meglecz,E., Petenian,F., Danchin,E., Coeur D′Acier,A., Rasplus,J.-Y. and Faure,E. (2004) High similarity between flanking regions of different microsatellites detected within each of two species of Lepidoptera: *Parnassius apollo* and *Euphydryas aurunia*. *Mol. Ecol.*, **13**, 1693–1700.
20. Savilahti,H., Rice,P.A. and Mizuuchi,K. (1995) The phage Mu transpososome core: DNA requirements for assembly and function. *EMBO J.*, **14**, 4893–4903.
21. Yuan,J.F., Beniac,D.R., Chaconas,G. and Ottensmeyer,F.P. (2005) 3D reconstruction of the Mu transposase and the Type 1 transpososome: a structural framework for Mu DNA transposition. *Genes Dev.*, **19**, 840–852.
22. Haapa,S., Taira,S., Heikkinen,E. and Savilahti,H. (1999) An efficient and accurate integration of mini-Mu transposons *in vitro*: a general methodology for functional genetic analysis and molecular biology applications. *Nucleic Acids Res.*, **27**, 2777–2784.
23. Haapa,S., Suomalainen,S., Eerikäinen,S., Airaksinen,M., Paulin,L. and Savilahti,H. (1999) An efficient DNA sequencing strategy based on the bacteriophage Mu *in vitro* transposition reaction. *Genome Res.*, **9**, 308–315.
24. Kekarainen,T., Savilahti,H. and Valkonen,J.P. (2002) Functional genomics on potato virus A: virus genome-wide map of sites essential for virus propagation. *Genome Res.*, **12**, 584–594.
25. Poussu,E., Vihinen,M., Paulin,L. and Savilahti,H. (2004) Probing the α-complementing domain of *E. coli* β-galactosidase with use of an insertional pentapeptide mutagenesis strategy based on Mu *in vitro* DNA transposition. *Proteins*, **54**, 681–692.
26. Poussu,E., Jäntti,J. and Savilahti,H. (2005) A gene truncation strategy generating N- and C-terminal deletion variants of proteins for functional studies: mapping of the Sec1p binding domain in yeast Mso1p by a Mu *in vitro* transposition-based approach. *Nucleic Acids Res.*, **33**, e104.
27. Vilen,H., Aalto,J.M., Kassinen,A., Paulin,L. and Savilahti,H. (2003) A direct transposon insertion tool for modification and functional analysis of viral genomes. *J. Virol.*, **77**, 123–134.
28. Sambrook,J., Fritsch,E.F. and Maniatis,T. (2001) *Molecular Cloning: A Laboratory Manual (3 volume set)*. Cold Spring Harbor Laboratory Press, New York.
29. Baker,T.A., Mizuuchi,M., Savilahti,H. and Mizuuchi,K. (1993) Division of labor among monomers within the Mu transposase tetramer. *Cell*, **74**, 723–733.
30. Li,W.H. and Sadler,L.A. (1991) Low nucleotide diversity in man. *Genetics*, **129**, 513–523.
31. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
32. Rozen,S. and Skaletsky,H. (2000) Primer3 on the WWW for general users and for biologist programmers. *Methods Mol. Biol.*, **132**, 365–386.
33. Pajunen,M.I., Pulliainen,A.T., Finne,J. and Savilahti,H. (2005) Generation of transposon insertion mutant libraries for Gram-positive bacteria by electroporation of phage Mu DNA transposition complexes. *Microbiology*, **151**, 1209–1218.
34. Haapa-Paananen,S., Rita,H. and Savilahti,H. (2002) DNA transposition of bacteriophage Mu. A quantitative analysis of target site selection *in vitro*. *J. Biol. Chem.*, **277**, 2843–2851.
35. Palo,J., Varvio,S.-L., Hanski,I. and Väinolä,R. (1995) Developing microsatellites markers for insect population structure: complex variation in a checkerspot butterfly. *Hereditas*, **123**, 295–300.
36. Sarhan,A. (2006) Isolation and characterization of five microsatellite loci in the Glanville fritillary butterfly. *Mol. Ecol. Notes*, **6**, 163–164.
37. Gilchrist,E. J. and Haughn,G. W (2005) TILLING without a plough: a new method with applications for reverse genetics. *Curr. Opin. Plant Biol.*, **8**, 211–215.
38. Oleykowski,C.A., Mullins,C.R.B., Godwin,A.K. and Yeung,A.T. (1998) Mutation detection using a novel plant endonuclease. *Nucleic Acid Res.*, **26**, 4597–4602.
39. Yeung,A.T., Hattangadi,D., Blakesley,L. and Nicolas,E. (2005) Enzymatic mutation detection technologies. *Biotechniques*, **38**, 749–758.
40. Garrity,P.A. and Wold,B.J. (1992) Effects of different DNA polymerases in ligation-mediated PCR: enhanced genomic sequencing and *in vivo* footprinting. *Proc. Natl. Acad. Sci. U.S.A.*, **89**, 1021–1025.
41. Vigneault,F. and Drouin,R. (2005) Optimal conditions and specific characteristics of Vent exo$^-$ DNA polymerase in ligation-mediated polymerase chain reaction protocols. *Biochem. Cell Biol.*, **83**, 147–165.