

## ORIGINAL ARTICLE

# Evolutionary dynamics of multilocus microsatellite arrangements in the genome of the butterfly *Bicyclus anynana*, with implications for other Lepidoptera

AE Van't Hof<sup>1</sup>, PM Brakefield<sup>1</sup>, IJ Saccheri<sup>2</sup> and BJ Zwaan<sup>1</sup><sup>1</sup>Department of Evolutionary Biology, Institute of Biology, Leiden University, Leiden, The Netherlands and <sup>2</sup>School of Biological Sciences, The Biosciences Building, University of Liverpool, Liverpool, UK

The sequences flanking microsatellites isolated from the butterfly *Bicyclus anynana* display high levels of similarity among different loci. We examined sequence data for evidence of the two mechanisms most likely to generate these similarities, namely recombination mediated events, such as unequal crossing over or gene conversion and through transposition of mobile elements (MEs). Many sequences contained tandemly arranged microsatellites, lending support to recombination as the multiplication mechanism. There is, however, also support for ME-mediated multiplication of microsatellites and their flanking sequences. Homology with a known Lepidopteran ME was found in *B. anynana* microsatellite regions, and polymorphic microsatellite markers with partial similarities in their flanking sequences were passed on to the next generation independently, indicating that they are not linked. Therefore, the rise

of these similarities appears to be mediated through both processes, either as an interaction between the two, or by each being responsible for part of the observations. A large proportion of microsatellites embedded in repetitive DNA is representative for most studied butterflies and moths, and a BLAST survey of the *B. anynana* sequences revealed four short microsatellite-associated sequences that were present in many species of Lepidoptera. The similarities usually start to deviate beyond these sequences, which suggests that they define the extremes of a repeated unit. Further study of these conserved sequences may help to understand the mechanism underlying the multiplication events, and answer the question of why these redundancies are predominantly found in this insect group.

*Heredity* (2007) **98**, 320–328. doi:10.1038/sj.hdy.6800944; published online 28 February 2007

**Keywords:** microsatellites; unequal crossing-over; mobile elements; holocentric chromosomes; null-alleles

## Introduction

Microsatellites, consisting of tandemly repeated units of 2–6 bp, have proved to be one of the most versatile molecular markers available owing to their high level of repeat number variation and widespread distribution in eukaryotic genomes. The classical model for their evolution proposes that the initial repeated motif seed arises through random base substitution, followed by stepwise mutation through replication slippage (reviewed by Ellegren, 2004). However, the expanding microsatellite database, particularly from Lepidoptera, suggests that other mechanisms play an important role in the genesis of microsatellites.

In Lepidoptera, microsatellites and their flanking sequences often possess features which are uncommon in most other taxa. These features have impeded development of microsatellite markers, as illustrated by the relative paucity of lepidopteran microsatellites

described in the literature (Nève and Megléc, 2000; Supplementary on-line appendix 1).

Only recently has the collective set of observations been recognized as a genetic phenomenon in itself rather than being treated as a methodological nuisance for obtaining an acceptable number of markers (Megléc *et al.*, 2004; Zhang, 2004). The four major features of Lepidopteran microsatellites that have been suggested as possible causes of these low yields of markers are: (i) low genomic frequency of microsatellites, (ii) low proportions of polymorphic vs monomorphic markers, (iii) unstable flanking sequences interfering with polymerase chain reaction (PCR) amplification and (iv) multiple occurrences of similar flanking sequences. The following sections treat these reported features in turn.

- (i) Frequencies of microsatellites in Lepidoptera have been described in *Parnassius mnemosyne* and *Bombyx mori* (Megléc and Solignac, 1998; Reddy *et al.*, 1999; Prasad *et al.*, 2005). These show an average CA-repeat occurrence every 97 kb in *P. mnemosyne* and 40 kb in *B. mori*, which is larger than the interval found in most other taxa, but not unusual, and not nearly enough to explain the differences in yields with other (insect) groups (Nève and Megléc, 2000).

Correspondence: Dr BJ Zwaan, Department of Evolutionary Biology, Institute of Biology, Leiden University, PO Box 9516, Leiden 2300RA, The Netherlands.

E-mail: b.j.zwaan@biology.leidenuniv.nl

Received 12 July 2006; revised 24 November 2006; accepted 22 December 2006; published online 28 February 2007

Moreover, enrichment techniques used in the more recent studies did not substantially improve genetic marker yields, implying that the relative scarcity of microsatellites is not the primary cause for the poor results.

- (ii) Where specified, the proportion of monomorphic loci is usually low in Lepidoptera, and never high enough to explain the low number of discriminating markers as can be seen in Supplementary on-line appendix 1.
- (iii) Heterozygote deficiency has been reported in a large proportion of markers in most Lepidoptera studies (Supplementary on-line appendix 1). This is primarily caused by the frequent occurrence of null alleles (Cassel, 2002; Jiggins *et al.*, 2005; Van't Hof *et al.*, 2005). There is substantial evidence that many null alleles in Lepidoptera are caused either by mutations in primer binding sites resulting in unsuccessful PCR, or by indels that produce alleles with PCR fragment sizes which fall outside the standard detection range (Palo *et al.*, 1995; Keyghobadi *et al.*, 1999; Reddy *et al.*, 1999; Flanagan *et al.*, 2002; Jiggins *et al.*, 2005). Therefore, this relatively high flanking sequence variability, that manifests itself as null alleles, is in part responsible for the low yields.
- (iv) The primary cause of the difficulties in obtaining markers, however, is not that flanking sequences differ too much for successful amplification as described above, but rather that these sequences at more than one locus are too much alike. This usually results in more than two different distinguishable PCR products, causing uninterpretable banding patterns (Palo *et al.*, 1995; Bogdanowicz *et al.*, 1997; Anthony *et al.*, 2001; Williams *et al.*, 2002; Ji *et al.*, 2003).

Our own data, based on several microsatellite-enriched libraries of the Afrotropical butterfly, *Bicyclus anynana* (Satyridae), are consistent with such unusual microsatellite characteristics (Van't Hof *et al.*, 2005). Thus, we found that most sequences surrounding microsatellites show similarities. Of these, we found those with similar sequences on both sides of the microsatellite and those where only one flank matches other sequences. These two categories of flanking sequence similarity have been named symmetrical and asymmetrical by Megléc *et al.* (2004) after finding analogous structures in two other butterfly species.

The present study focuses on the origins of the multiplications that have led to these multicopy sequences, and on why this process is so widespread in Lepidoptera. We first consider the possibility that asymmetrical sequences might in fact be artifacts, representing chimeric PCR products formed during the enrichment PCR step (Pääbo *et al.*, 1990).

Secondly, we focus on the mechanisms through which multicopy DNA arises and how they are involved in *B. anynana* microsatellites. The two main pathways are by means of transposition of mobile elements (MEs) and by recombination. We surveyed the data set for tandemly repeated patterns as would be the case after unequal crossing over or gene conversion, and also screened it for ME characteristics such as direct or inverted repeats and for similarities with sequence data for known MEs. Furthermore, we examined whether the microsatellites

co-migrate within their surrounding sequences or whether they were formed from proto-microsatellites after the multiplication event, as is the case in mini-me's in *Drosophila* (Wilder and Hollocher, 2001), primate Alu elements (Arcot *et al.*, 1995), and in introns of human and desert locust (*Schistocerca gregaria*) FABP genes (Wu *et al.*, 2001).

Finally, we consider our data in a broader perspective by making comparisons to other species with a particular emphasis on the Lepidoptera. We thus aim to find clues about a unitary mechanism, and to find out why these phenomena are mainly reported from butterflies and moths.

## Materials and methods

### DNA extraction, library construction and sequencing

The source material for all analysed sequences is DNA extracted from thorax and head of a single butterfly using a standard phenol-CIA protocol as described in Van't Hof *et al.* (2005). A female was used to incorporate both the W and Z chromosomes. Enrichment for CA, GA, AAT, ATG, GAA and TACA motifs was performed by Genetic Identification Services (GIS, <http://www.genetic-id-services.com>; Chatsworth, CA, USA) using *Hind*III restriction and adapters, and a single round of enrichment with biotinylated microsatellite sequences as capture molecules. Positive DNA fragments of 350–700 bp were cloned in pUC19. The libraries were transformed into JM109 (Promega, Madison, WI, USA), followed by blue–white screening. Positive clones were grown in 200  $\mu$ l LB with 100  $\mu$ g/ml ampicillin and miniprepped using the Qiaprep spin miniprep kit (Qiagen, Valencia, CA, USA). Sequencing was outsourced to commercial facilities. The numbers of sequenced clones per library are given in Table 1.

### Detection of intra-specific similarities

Similarities within this data set were detected by comparing the sequences from all libraries with each other by means of 'all against all' standalone nucleotide–nucleotide BLAST (BLASTN) (Altschul *et al.*, 1997) and then manually fine-aligning where needed using BIOEDIT (Hall, 1999). The length threshold for considering sequence homologues was set to 40 bp in addition. Shorter homologues with adjoining microsatellites that were omitted by BLASTN owing to their repetitive nature were included.

### Detection of inter-specific homologues

Homologies between our data and sequences submitted to GenBank were surveyed with online BLASTN using default settings. Distinction between hits that occur by chance and true 'common origin' data is not fully represented in the 'Blast Score' as it does not compensate for the differences in available sequences per species. Therefore, we used a threshold of 50 to include hits from large-scale genome surveys, and a threshold of 40 for species with under-represented sequence data resources. The hits matching these criteria were then manually realigned with BIOEDIT for two reasons. First of all, repeat structures are not included in the BLASTN output whereas the detected match often continued into a shared microsatellite or even beyond, and secondly,

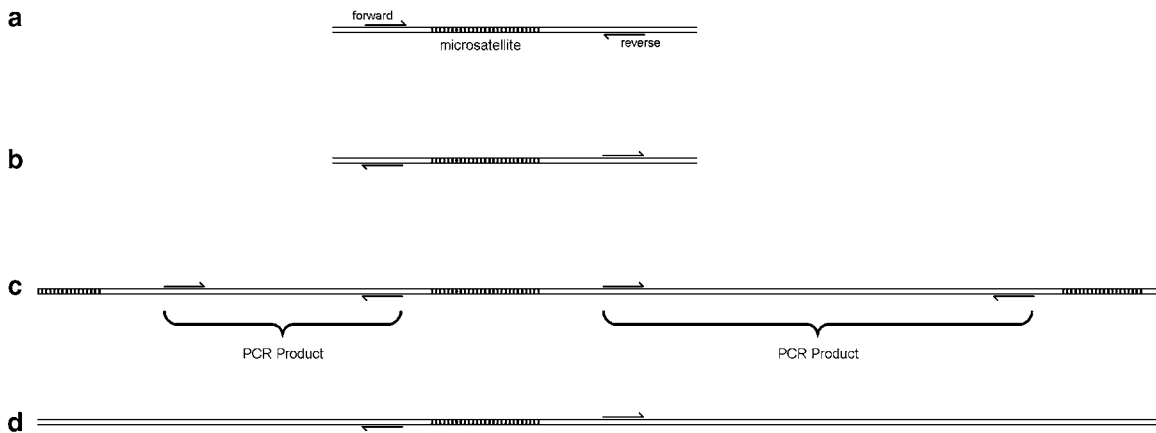
**Table 1** Properties of the sequences extracted from the six enriched libraries

Redundancy and repeat type categories	Libraries						Total characteristics [shared characteristics]
	CA	GA	AAT	ATG	GAA	TACA	
SC microsat	24	5 <sup>a</sup>	3	6	0	3	41 <sup>a</sup> [1]
MC microsat (# of groups)	117 (12) <sup>a</sup>	3 (1)	13 (4)	34 (4)	8 (3)	9 (2)	184 <sup>a</sup> [4]
SC minisat	7	1 <sup>a</sup>	4	2	2	1	17 <sup>a</sup> [1]
SC minisat with microsat.	9	2	0	1	0	1	13
MC minisat (# of groups)	12 (4) <sup>a</sup>	0	2 (1)	0	0	0	14 <sup>a</sup> [4]
MC minisat with microsat (# of groups)	2 (1)	0	0	2 (1)	1	1	6
No tandem repeats (of which MC)	6 (1)	0	1 (1)	8 (5)	3 (2)	1 (0)	19 (9)
Total characteristics (total clones)	177 (173) <sup>a</sup>	11 (10) <sup>a</sup>	23	53	14	16	294 (289) <sup>a</sup>

Abbreviations: SC, single copy, MC, multicopy.

For the three different MC classes, the numbers of homologous groups are given for the intra-library homologies '(# of groups)'. The category 'MC microsat' is composed of symmetrical, asymmetrical and partial homologies. 'No tandem repeats' consists of SC- and MC clones without microsatellite or minisatellite structures.

<sup>a</sup>Some sequences contain both a microsatellite and a minisatellite (not to be confused with a microsatellite inside a minisatellite) and are, therefore, included twice in the statistics. For that reason, the table states both 'total characteristics' and 'total clones'. The sum of '(shared characteristics)' divided by 2 ( $10/2=5$ ), subtracted from 'Total characteristics' provides the total number of clones ( $294-5=289$ ).



**Figure 1** Response of PCR amplification to different microsatellite-flank arrangements. (a) Example of normal microsatellite primer design with a forward and a reverse primer on either side of the repeat, initiating polymerization directed towards each other. (b) Primer design for this particular experiment with primers oriented away from the microsatellite and, more importantly, away from each other. (c and d) The two possible scenarios; (c) tandem arrangement with a relatively short distance between the units, resulting in exponential amplification, or (d) no tandem arrangement, or large repeat units with too distant primer recognition sites for successful amplification.

many obvious homologies surrounding the returned sequence match were not reported by BLASTN.

Sequence regions that were reported from multiple species were aligned with BIOEDIT to construct a consensus sequence. Subsequently, this sequence was re-analysed with online BLASTN, followed by an update of the consensus based on the additional hits. This process was repeated until no more new hits occurred.

#### Experiment I: confirming the presence of specific sequences in genomic DNA

To test whether the different combinations of flanking sequences were an artifact caused by enrichment procedures, or in fact occur in the observed association in the butterfly genome, we designed primers with OLIGO version 6 (Rychlik, 2000) to amplify 15 different combinations of symmetrical and asymmetrical sequence clusters in the ATG library. Product was detected with ethidium bromide-stained 1% agarose gel. PCR was performed in 10  $\mu$ l, containing 5  $\mu$ l 2  $\times$  Reddymix 1.5

(ABgene), 0.33  $\mu$ M of each primer, with 1  $\mu$ l second elution DNeasy-tissue (Qiagen) extracted thorax as template. Thermal cycle was 3 min 95°C; 30 cycles of 30 s 94°C, 30 s  $T_a$ , 45 s 72°C, followed by 30 min at 72°C.  $T_a$  was 50°C for all but primer-pair 9 (BA-ATG244), where  $T_a=47^\circ\text{C}$ . The primer sequences are listed in Supplementary on-line appendix 2.

#### Experiment II: exploration of the spatial organization of common sequences

PCR primers were designed with an outward orientation instead of inward on both ends of the cloned insert (i.e. primers amplifying away from the microsatellite instead of towards it, as in inverse PCR). They were based on the consensus sequences of six symmetrical (microsatellite flanking sequence) groups (AAT group 1, ATG group 2A upstr.A-dstr.A, ATG group 2B upstr.F-dstr.A, CA groups 1–3). This arrangement of primers will only result in amplification if the complementary primer is within range (see Figure 1). PCR was performed as in experi-

ment I, but with a 55°C  $T_a$  for ATG group 2A upstr. A-dstr.A, which is 5°C above the recommended  $T_a$  to reduce the chance of non-specific priming. Amplification products were detected with ethidium bromide-stained 1% agarose gel. The primer sequences are given in Supplementary on-line appendix 3.

A band of approximately 275 bp that consisted of two merged amplicons produced with the ATG group 2A upstr.A-dstr. A primer combination was excised and purified with Qiaquick gel extraction kit (Qiagen) and inserted into a cloning vector using the pGEM-T system (Promega). Transformation, cloning and sequencing were performed as described above for the enriched library.

## Results

### Sequence similarities within the *B. anynana* libraries

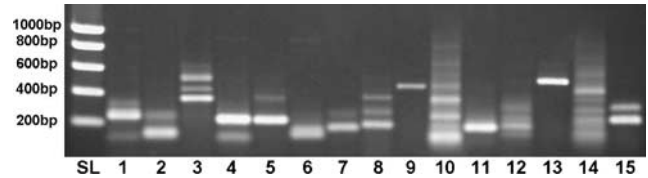
Most sequences from the *B. anynana* enriched libraries showed typical Lepidopteran microsatellite characteristics, such as symmetrical and asymmetrical flanking regions surrounding the repeat structure. These multi-copy sequences were found in all of the six libraries and their details are summarized in Table 1. The standalone 'all against all' BLASTN revealed that sequences are not only associated within the different enriched libraries, but also frequently between them. Compound microsatellites selected by multiple enrichment probes make up just a small fraction of these intra-library links. The proportion of clones that show no similarity is 80 out of 289, which is an overestimate, as large numbers of redundant clones were filtered out before sequencing (Van't Hof et al., 2005). Sequence data have been deposited with the EMBL/GenBank Data Libraries under accession nos. AY785060, AY785062, AY785064, AY785071, AY785080, AY785081, DQ225274-DQ225304, EF114667-EF114669.

### Confirmation of the presence of cloned sequences in genomic DNA

The PCR amplification of different asymmetrical combinations gave robust amplification products in each of the 15 different upstream-downstream primer combinations (Figure 2). This showed that the observed data are not an enrichment artifact, but these asymmetrical structures actually occur as contiguous sequences in the *B. anynana* genome. Most of the PCR products showed more than one distinct band, indicative of multiple copies with a variable distance between the primer binding sites.

### Relative orientation of common sequences

A sequence family from the ATG library is represented as a schematic alignment in Figure 3 in to provide an example of the similarity patterns. The ATG2 sequence family consists of two subgroups that are linked together by sequences that possess characteristics of both clusters (BA-ATG244 and BA-ATG1). Subgroup 2A is defined by a 60 bp sequence directly adjacent (upstream) to the ATG<sub>n</sub> repeat (2A upstr. A), and subgroup 2B is characterized by a 31 bp sequence immediately beyond a common CAT<sub>n</sub> repeat (2B downstr. A). The relative positions of the different sequence regions are designated by (i) the alignment subgroup (2A or 2B), (ii) their position upstream/downstream (u/d) relative to the



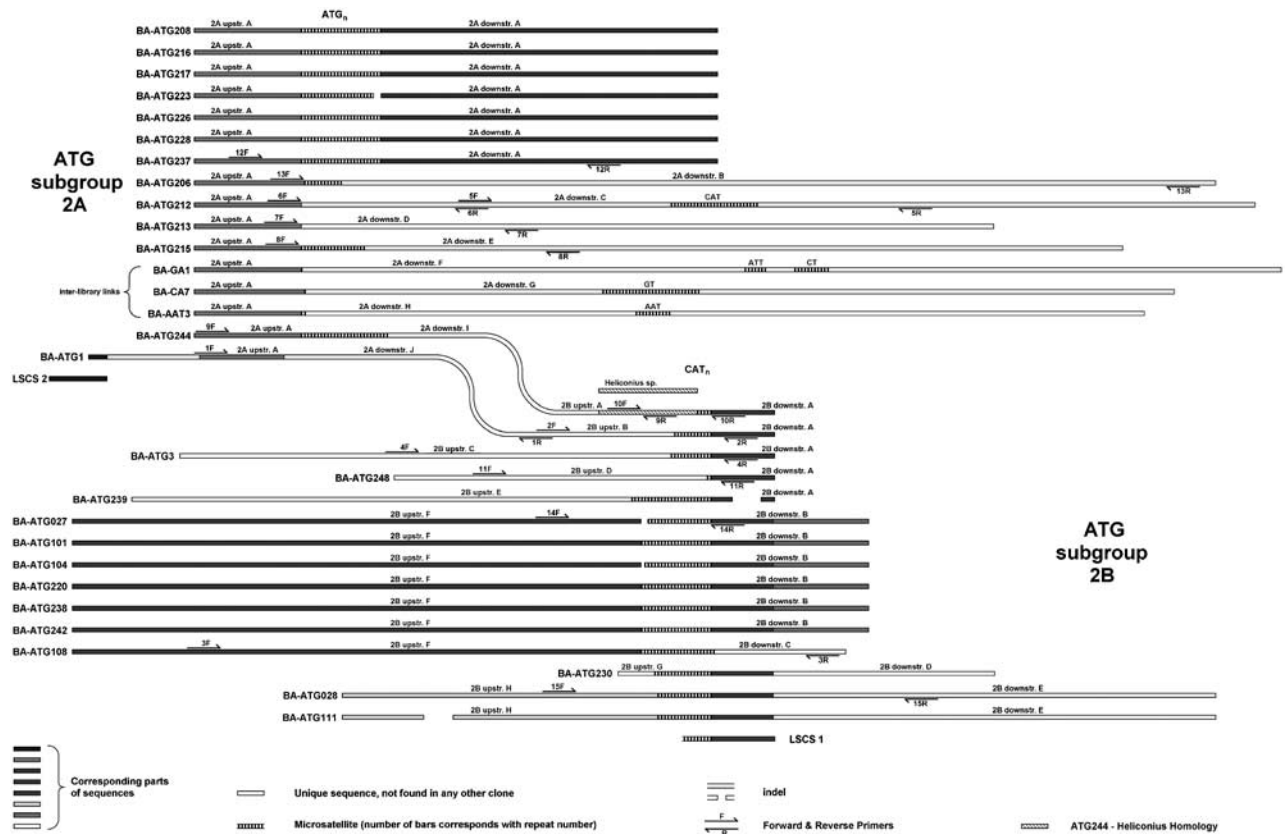
**Figure 2** PCR product from 15 different primer pair combinations designed to test sequence associations found in the ATG library. Lane numbers correspond to the following primer combinations (see Figure 3 for primer locations): SL = Eurogentec Smartladder; 1 = BA-ATG1 subgroup 2A upstream A and downstream J (2A-uA-dJ); 2 = BA-ATG1/2B-uB-dA; 3 = BA-ATG108/2B-uF-dC; 4 = BA-ATG3/2B-uC-dA; 5 = BA-ATG212/ single copy microsatellite region; 6 = BA-ATG212/2A-uA-dC; 7 = BA-ATG213/2A-uA-dD; 8 = BA-ATG215/2A-uA-dE; 9 = BA-ATG244/2A-uA-dI; 10 = BA-ATG244/2B-uA-dA; 11 = BA-ATG248/2B-uD-dA; 12 = consensus 2A-uA-dA; 13 = consensus 2A-uA-dB; 14 = consensus 2B-uF-dD; 15 = consensus 2B-uH-dE.

aligned microsatellites and (iii) by their class of similarity within each subgroup (A–J). Two clusters, 2A-uA-dA and 2B-uF-dB, are examples of symmetrical associations, possessing similarities on both sides of the microsatellites. Both subgroups also have many asymmetrical associations with some flanks overrepresented, rather than a random mixture of upstream-downstream combinations (e.g. BA-ATG206, 212, 213 and 215). The prevalence of one type of flank on one side and variation on the other side of the microsatellite is a characteristic of most other asymmetrical groups that were found in *B. anynana*. Asymmetrical inter-library alignments are represented in Figure 3 by BA-GA1, BA-CA7 and BA-AAT3. They match with 2A-uA, followed by an ATG<sub>1</sub> or ATG<sub>2</sub> in line with the ATG<sub>n</sub> site.

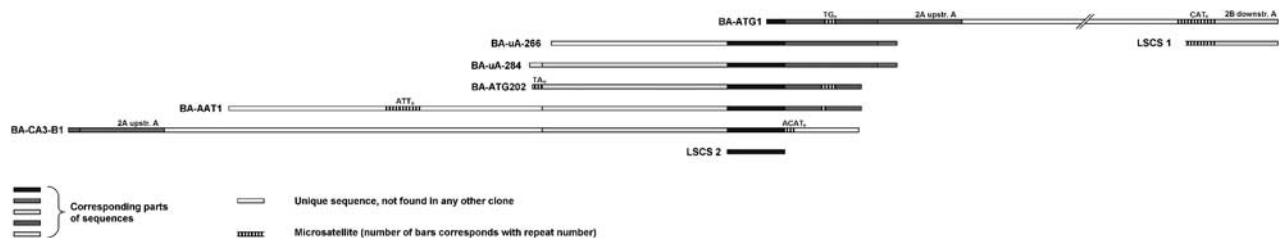
The two main aligned microsatellites in Figure 3 differ markedly in repeat numbers with 0 to 26 repeats in 2A and three to 29 in 2B. Additional microsatellites present in some '2A' sequences appear to be unrelated to the aligned ATG<sub>n</sub>, and consist of different repeat types. These sequences often align partially or asymmetrically to other sequences or groups of sequences either within or between libraries (not shown in Figure 3).

The experiment to explore sequences surrounding some of the sequence families gave a positive amplification result in three of the six combinations (ATG-2A-uA-dA, ATG-2B-uF-dA, CA group 3). This implies that some common sequences are repeated relatively closely beyond the known sequence.

The ATG-2A-uA-dA band that was sequenced from these PCR products consists of a 266 and a 284 bp fragment (BA-uA-266 and BA-uA-284). They both match with the upstream-A flank, including the BA-ATG1 extension (Figure 3), but shared little more than the primer sequence with the ATG-2A-dA region. The BA-uA-266 and BA-uA-284 sequences form a link between a sequence cluster consisting of BA-ATG202, BA-AAT1 and BA-CA3-B1 plus the upstream part of the ATG2A subgroup. The schematic alignment of these sequences is presented in Figure 4, which has partial overlap with Figure 3. The BA-uA-266 and BA-uA-284 sequences are nearly identical for about half their length, but lose their similarity immediately after a 35 bp non-random sequence that is associated with multicopy microsatellites in many Lepidoptera species. This sequence, designated LSCS2, will be discussed in detail below.



**Figure 3** Schematic representation of the alignment of ATG group 2 sequences, showing two subgroups (2A and 2B) linked together by two sequences possessing characteristics of both. The majority are intra-library links, grouped to symmetrical sequence families with variable microsatellites, and asymmetrical alignments. Inter-library links are shown in subgroup ATG 2A and inter-specific hits are represented in subgroup ATG 2B by the Lepidoptera Specific Core Sequence LSCS1 (see inter-specific comparison section in Results) and a *Heliconius* sequence. The arrows represent the location of the primers used in the control experiment to verify the existence of several upstream-downstream combinations, with forward (F) and reverse (R) primer numbers corresponding to the lane numbers in Figure 2.



**Figure 4** Schematic alignment of sequences upstream of subgroup ATG2A, showing the full LSCS 2 alignment. There is partial overlap with Figure 3.

The ATG-2A-uA sequence that characterizes subgroup ATG 2A recurs further upstream in the BA-CA3-B1 sequence (Figure 4). Furthermore, this group of sequences incorporates a microsatellite that is variable in repeat number, but whose variability does not alter the overall length of the sequence (i.e. caused by base substitutions rather than by means of DNA replication slippage). This could either represent the different stages of a developing proto-microsatellite or a microsatellite in decay.

#### Inter-specific comparison with *B. anynana* microsatellite sequences

The online BLASTN comparison of the *B. anynana* sequences resulted in hits with nine butterflies, 23 moths,

one Coleoptera, two Diptera and two Hymenoptera (the species list is available in Supplementary on-line appendix 4). Four distinct Lepidoptera-specific core sequences (LSCS), nearly exclusively matching a wide range of Lepidoptera species, were identified from these BLAST hits. They are generally situated next to a microsatellite, and usually define the position where similar regions start to differ in sequence.

LSCS1 is a 38 bp sequence that corresponds to the ATG2B-dA sequence that is aligned in Figures 3 and 4. A BLAST search of this core sequence results in over 40 hits within 15 Lepidoptera species and one Coleoptera species (*Diabrotica virgifera*). With one exception, they all have a microsatellite in the same position as the CAT<sub>n</sub> region in *B. anynana*. In addition to the predominant CAT<sub>n</sub> repeats in these BLAST hits, several of these

sequences also contain ATT<sub>n</sub>, CCAT<sub>n</sub>, CAAT<sub>n</sub> or CA<sub>n</sub> arrays. The LSCS1 in *D. virgifera* is sandwiched in between two microsatellites (CAT<sub>n</sub> and CA<sub>n</sub>). The 35 bp LSCS2 matches the common sequence in the aligned cluster shown in Figure 4, and also aligns with the extreme end of BA-ATG1 (Figure 3). This core sequence is present in 13 deposited sequences from eight Lepidoptera species. In contrast to the other three LSCSs, this sequence is not typically bordered by a microsatellite, although there is a small microsatellite immediately beyond it in BA-CA3-B1. The 150 bp LSCS3 was detected in 11 Lepidoptera species, based on the BA-TACA105 BLAST hits. It spans both flanks of a common CAAA<sub>n</sub> microsatellite and is associated with retrotransposons in *Bombyx mandarina* (GenBank no. AB055223), *B. mori* (GenBank no. AB032718) and *Antheraea mylitta* (GenBank no. AF530471). The LSCS4, identified from BA-TACA112, consists of an 85 bp sequence and was found in six Lepidoptera species, usually bordered by a microsatellite. The sequences of the four LSCSs and the alignments with their BLAST hits are presented in on-line Supplementary appendices 5–9.

Besides the four core sequences that were present in many Lepidoptera species, there were also a number of more solitary hits, but still predominantly from Lepidoptera and often associated with microsatellites. One of these inter-specific links is represented in Figure 3 by *Heliconius cydno* and *H. melpomene* microsatellite flanks corresponding to part of the BA-ATG244 sequence.

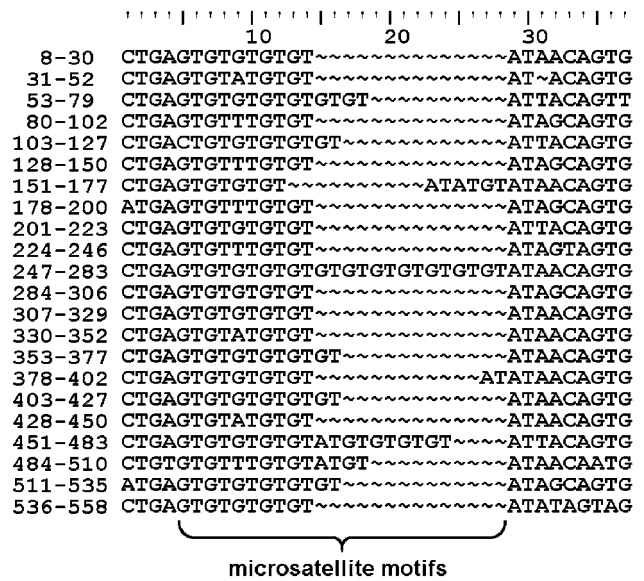
### Minisatellite structures

In addition to the microsatellites, 15% of the clones contained minisatellites with repeat units ranging from 14 to 55 bp, either with or without a microsatellite incorporated within each unit. Most of the microsatellites embedded in minisatellite units showed repeat number variation, which is possibly (but not necessarily) caused by slipped strand mispairing (relatively frequently occurring mutations adding or removing a repeat unit) as is the case in solitary microsatellites (Figure 5).

Many of the minisatellites could be grouped together in gene families in the same way as described above for the multicopy microsatellites. The different representatives of each family show variation in number of repeat units, composition of the units and of their flanking sequences. An overview of the numbers of clones containing the different minisatellite characteristics can be found in Table 1. The 10 bp Jeffreys core sequence (GGCAGGANG) (Jeffreys *et al.*, 1985) was found as a 9 out of 10 base match and a 100% match in the repeat units of BA-CA4-C1 and BA-AAT2-B11, respectively.

### Discussion

In Goldstein and Schlötterer (1999), the flanking region is described as ‘The single-copy DNA sequence immediately upstream and downstream of a microsatellite locus that allows the design of specific primers that preferentially amplify the target microsatellite’. The *B. anynana* data set presented here suggests that this definition cannot be universally applied, because most microsatellites in this species are located within repetitive DNA. This appears to be a general characteristic of Lepidoptera (Supplementary on-line appendix 1), and has also been found in some other insects, such as Coleoptera



**Figure 5** Internal alignment of a 551 bp stretch of BA-CA1-G4, showing 22 minisatellite units with incorporated variable microsatellites (GT<sub>4</sub>–GT<sub>12</sub>).

(Liewlaksaneeyanawin *et al.*, 2001; N. Margraf pers. comm) and Diptera (Fagerberg *et al.*, 2001; Wilder and Hollocher, 2001). Apart from observations in insects, microsatellites associated with repetitive DNA have also been reported in vertebrates (Alexander *et al.*, 1995; Arcot *et al.*, 1995; Band and Ron, 1996; Nadir *et al.*, 1996; Gallagher *et al.*, 1999) and in plants (Ramsay *et al.*, 1999; Temnykh *et al.*, 2001; Tero *et al.*, 2006).

The possibility that multiple variants of a certain locus were incorporated into the genomic library by means of chimeric reassociation during the PCR-based enrichment (Pääbo *et al.*, 1990) was dismissed by the successful genomic PCR amplification of 15 different repetitive DNA sequences. We usually found amplicons of different sizes per amplification, indicating that they originate from multiple loci (Figure 2). A similar experiment was performed by Tero *et al.* (2006), who found that 82.1% of the tested primer combinations confirmed that the sequences derived from their genomic library were contiguous in *Silene tatarica*, and sequencing of amplification products of different sizes revealed that they represent heterogeneous loci.

Another indication that the sequences obtained from the *B. anynana* library are contiguous is given by the fact that a number of sequences with similar regions can be amplified uniquely and serve as polymorphic microsatellite markers as long as the primers target unique parts of these sequences (BA-GA1, BA-CA7, BA-AAT3, BA-ATG1 and BA-ATG3, all represented in Figure 3).

Enrichment procedures may, however, have a bias towards certain sequences other than the repeat itself. For instance, the BA-ATG213 sequence that belongs to the ATG2 family was included in the library in spite of not containing a microsatellite.

### Repeat unit definition

The two main mechanisms for multiplication of DNA sequences are by means of transposition of MEs that

have the ability to incorporate copies of themselves elsewhere in the genome, or even in other individuals, and recombination-related events, such as unequal crossing over (UCO) and gene conversion that result in tandemly arranged homologues.

One limitation of the material studied here to distinguish between these two possibilities is that it is not always clear what defines the extremes of a repeated unit. Inserts were selected in the 350–700 bp range, whereas many MEs and recombination products are larger. There are, however, two common structures where similar sequences start to differ. First, asymmetrical sequences are by definition identical on one side of the microsatellite and different on the other side. Second, there are the LSCS structures that usually define the start of sequence divergence.

### Mobile elements

There is support in our data for the hypothesis that MEs are responsible for the abundance of similarity surrounding microsatellites. The BA-TACA105 derived LSCS3 fully matched to Lepidopteran retrotransposons of *Bombyx mori*, *B. mandarina* and *Antheraea mylitta*. It is possible that the other three LSCSs are structural units of MEs as well. The fact that LSCS1 and LSCS2 are present in a single sequence (BA-ATG1) would indicate that they define different parts of the same mobile element.

MEs usually have specific characteristics such as inverted or direct repeats at their extremes, or poly-A tracts (for an overview see Berg and Howe, 1989). A small number of short direct and inverted repeats were found in *B. anynana*, and 25 sequences contained a poly-A of 10 or more base pairs.

Another observation in *B. anynana* that supports MEs rather than recombination is the independent inheritance of asymmetrical loci in an F2 cross, indicating that the microsatellites in question are not closely linked (Van't Hof et al., 2005).

Examples of ME-associated microsatellites in other Lepidoptera species are those in the very common *Bombyx mori* BM1 elements, which are 'surrounded by short direct repeats (2–6 bp)' (Eickbush, 1995) and the similarities between *Parnassius* microsatellite clones and a *Drosophila* retrotransposable element and a human retrovirus (Megléczy et al., 2004).

At odds with the involvement of proto-microsatellite containing MEs (Wilder and Hollocher, 2001) are some very distinct polymorphisms that interrupt the microsatellites in *B. anynana*. They manifest themselves in different loci or repeat units (e.g. CA group 2, Supplementary on-line appendix 10), indicating that the microsatellites must have been present before the multiplication event, and hitchhiked in conjunction with the flanking sequences.

### Recombination as cause for repetitive sequences

There is also support for the involvement of recombination as a mechanism for part of the observed repetitive sequences from the present data set. Minisatellites are generated through recombination, and each minisatellite unit of a microsatellite-containing minisatellite can be described as a microsatellite with flanking regions, as in a solitary microsatellite, only with much shorter flanks. On a larger scale, the BA-CA3-E3 sequence shows two

tandemly arranged units of approximately 100 bp each, which both include a CA<sub>n</sub> repeat (CA<sub>9</sub> and CA<sub>13</sub>, respectively), and can also be defined as microsatellites with flanking sequences. When similar microsatellite-containing repeated units become much larger (i.e. larger than the cloned insert) it is impossible to detect their higher order repetitive nature. It is therefore possible that part of the repetitiveness is comparable to the microsatellite-containing minisatellites, but with a much larger unit size. The BA-CA3-B1 sequence indirectly positions the ATG2A-uA sequence upstream of the main ATG2A-uA alignment (Figure 4), which may represent tandem arrangement. The fact that the alignment ends after the ACAT<sub>n</sub> microsatellite in this sequence could be owing to an indel as described below, thus it is not unlikely that the ATG2A-uA sequence actually recurs downstream of this sequence.

The asymmetrical sequence arrangements fit perfectly within the description of UCO (i.e., where a chiasma occurs at two imperfectly aligned microsatellites with shared repeat units, leaving two new upstream–downstream combinations) (Megléczy et al., 2004). There are, however, some discrepancies. One of the features in Lepidoptera microsatellites is that they often possess indels of various sizes directly adjacent to the microsatellite (Reddy et al., 1999; Flanagan et al., 2002). If such an indel is too large to find a match within a sequence family, it may be misinterpreted as a completely different flank. For example, the BA-CA3-E11 clone, belonging to CA group 2, contains a 173 bp deletion immediately after the microsatellite, and rejoins at the end of the main alignment with a perfect match of 35 bp (Supplementary on-line appendix 10). Had the deletion been 35 or more bases larger, no matching sequence would have been found and it might have been wrongly attributed to misaligned-microsatellite UCO. The fact that there are instances where indels form an alternative explanation for the observed asymmetries does not however rule out recombination as a contributory mechanism for repetitiveness altogether.

### Lepidoptera-specific homologues

The comparison of *B. anynana* clones with GenBank resulted in a large number of hits that were very strongly biased towards butterflies and moths. One could argue that it is not surprising to BLAST Lepidoptera material and get Lepidoptera hits in return. The issue here, however, is that some regions seem to be very widely conserved in Lepidoptera, and, more importantly, they are associated with the very phenomena we are exploring, namely multicopy microsatellite flanking regions in Lepidoptera. It seems, therefore, that there is a shared mechanism involved in the Lepidoptera that is reflected in the conservation of certain sequences. In particular, the four LSCSs seem to be so frequent and widely distributed in this group that they may be key sequences for further investigation of these issues.

### Sister chromatid association in Lepidoptera

The impression that the patterns described are peculiar to Lepidoptera raises the question of what might distinguish them from other groups. One uncommon feature present in all Lepidoptera is their holocentric chromosome arrangement, where chromatids assemble

over their entire length instead of being joined at a centromere. Depletion of KLP-19, an essential microtubule motor, caused misalignment of holocentric kinetochores in the cabbage moth, *Mamestra brassicae* (Mandrioli *et al.*, 2003). This suggests a direct link between holocentric chromosomes in Lepidoptera and UCO. However, a survey of other species with holocentric chromosomes, including *Caenorhabditis elegans*, species of Hemiptera and certain plants did not reveal similar microsatellite flank redundancies, whereas other species that possessed them, such as some Coleoptera and Diptera, have centromere-associated chromosomes.

#### Overrepresentation of multicopy microsatellites vs under-representation of unique microsatellites

The low ratio of single copy to multicopy microsatellites from various studies on Lepidoptera has generally been interpreted as indicating high frequencies of the latter, relative to other taxa. An alternative, or complementary, interpretation is that single-copy microsatellites are scarce in Lepidoptera. This may also be reflected in the large number of null alleles reported in Lepidoptera, as if there are too few alternatives to these suboptimal microsatellite loci, they are more likely to be utilized and published. Prasad *et al.* (2005), interpreted the data cited in the Introduction as indicating that microsatellite densities are not unusually low in *Bombyx mori*; however, in this study the microsatellite densities obtained from more than 4400 *in silico* detected loci (total density of one locus per 6.4kb) are not separately specified as single copy and multicopy loci, which makes it difficult to determine whether multicopy microsatellites are unusually abundant or unique microsatellites scarce.

#### Conclusion

Our exploration of different hypotheses that may explain these unusual observations provided no clear-cut mechanism, as there is support for both recombination and MEs being implicated in the multiplication events. Therefore, a combination of both explains our observations best. The question remains as to whether we are dealing with two separate processes, that both lead to redundancy, or if it is an integrated mechanism.

Analysis of the repetitive microsatellite characteristics in *B. anynana* and other Lepidoptera species revealed a number of Lepidoptera-specific patterns that provides a basis for further research on this subject. The four core sequences appear to hold valuable information and may serve as a starting point for further investigations (e.g. *in situ* hybridization), leading to a better understanding of the mechanisms involved, and possibly in defining a new type of Lepidopteran mobile element. These findings may not only lead to a more complete knowledge of micro- and minisatellites in Lepidoptera, but may have general implications for understanding VNTR dynamics.

#### Acknowledgements

We thank two anonymous referees and the subject editor for comments on a previous version of this paper.

#### References

- Alexander L, Rohrer G, Beattie C (1995). Porcine SINE-associated microsatellite markers: evidence for new artiodactyl SINEs. *Mamm Genome* 6: 464–468.
- Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W *et al.* (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl Acid Res* 25: 3389–3402.
- Anthony N, Gelembiuk G, Raterman D, Nice C, Ffrench-Constant R (2001). Isolation and characterization of microsatellite markers from the endangered Karner blue butterfly *Lycaeides melissa samuelis* (Lepidoptera). *Hereditas* 134: 271–273.
- Arcot S, Wang Z, Weber J, Deininger P, Batzer M (1995). Alu repeats – a source for the genesis of primate microsatellites. *Genomics* 29: 136–144.
- Band M, Ron M (1996). Creation of a SINE enriched library for the isolation of polymorphic (AGC)<sub>n</sub> microsatellite markers in the bovine genome. *Anim Genet* 27: 243–248.
- Berg DE, Howe MM (1989). *Mobile DNA*. American Society for Microbiology: Washington, DC.
- Bogdanowicz SM, Mastro VC, Prasher DC, Harrison RG (1997). Microsatellite DNA variation among Asian and North American gypsy moths (Lepidoptera: Lymantriidae). *Ann Ent Soc Am* 90: 768–775.
- Cassel A (2002). Characterization of microsatellite loci in *Coenonympha hero* (Lepidoptera: Nymphalidae). *Mol Ecol Notes* 2: 566–568.
- Eickbush TH (1995). Mobile elements of lepidopteran genomes. In: Goldsmith MR, Wilkins AS (eds). *Molecular Model Systems in the Lepidoptera*. Cambridge University Press. pp 77–105.
- Ellegren H (2004). Microsatellites: simple sequences with complex evolution. *Nat Rev Genet* 5: 435–445.
- Fagerberg AJ, Fulton RE, Black WC (2001). Microsatellite loci are not abundant in all arthropod genomes: analyses in the hard tick, *Ixodes scapularis* and the yellow fever mosquito, *Aedes aegypti*. *Insect Mol Biol* 10: 225–236.
- Flanagan NS, Blum MJ, Davison A, Alamo M, Albarrán R, Faulhaber K *et al.* (2002). Characterization of microsatellite loci in neotropical *Heliconius* butterflies. *Mol Ecol Notes* 2: 398–401.
- Gallagher PC, Lear TL, Coogle LD, Bailey E (1999). Two SINE families associated with equine microsatellite loci. *Mamm Genome* 10: 140–144.
- Goldstein DB, Schlötterer C (1999). *Microsatellites; Evolution and Applications*. Oxford University Press Inc.: New York.
- Hall T (1999). BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucl Acids Symp Ser* 41: 95–98.
- Jeffreys AJ, Wilson V, Thein SL (1985). Hypervariable ‘minisatellite’ regions in human DNA. *Nature* 314: 67–73.
- Ji YJ, Zhang DX, Hewitt GM, Kang L, Li DM (2003). Polymorphic microsatellite loci for the cotton bollworm *Helicoverpa armigera* (Lepidoptera: Noctuidae) and some remarks on their isolation. *Mol Ecol Notes* 3: 102–104.
- Jiggins CD, Mavarez J, Beltrán M, McMillan WO, Johnston JS, Bermingham E (2005). A genetic linkage map of the mimetic butterfly, *Heliconius melpomene*. *Genetics* 171: 557–570.
- Keyghobadi N, Roland J, Strobeck C (1999). Influence of landscape on the population structure of the alpine butterfly *Parnassius smintheus* (Papilionidae). *Mol Ecol* 8: 1482–1495.
- Liewlaksaneeyanawin C, Ritland CE, Newton CH, El-Kassaby YA (2001). Characterization of microsatellite loci in white pine weevil (*Pissodes strobi*). *Mol Ecol Notes* 1: 248–249.
- Mandrioli M, Manicardi GC, Marec F (2003). Cytogenetic and molecular characterization of the MBSAT1 satellite DNA in holokinetic chromosomes of the cabbage moth, *Mamestra brassicae* (Lepidoptera). *Chromosome Res* 11: 51–56.
- Megléc E, Petenian F, Danchin E, D’Acier AC, Rasplus JY, Faure E (2004). High similarity between flanking regions of

- different microsatellites detected within each of two species of Lepidoptera: *Parnassius apollo* and *Euphydryas aurinia*. *Mol Ecol* **13**: 1693–1700.
- Megléc E, Solignac M (1998). Microsatellite loci for *Parnassius mnemosyne* (Lepidoptera). *Hereditas* **128**: 179–180.
- Nadir E, Margalit H, Gallily T, Ben-Sasson SA (1996). Microsatellite spreading in the human genome: evolutionary mechanisms and structural implications. *Proc Natl Acad Sci USA* **93**: 6470–6475.
- Nève G, Megléc E (2000). Microsatellite frequencies in different taxa. *TREE* **15**: 376–377.
- Pääbo S, Irwin DM, Wilson AC (1990). DNA damage promotes jumping between templates during enzymatic amplification. *J Biol Chem* **265**: 4718–4721.
- Palo J, Varvio S, Hanski I, Väinölä R (1995). Developing microsatellite markers for insect population structure: complex variation in a checkerspot butterfly. *Hereditas* **123**: 295–300.
- Prasad MD, Muthulakshmi M, Madhu M, Archak S, Mita K, Nagaraju J (2005). Survey and analysis of microsatellites in the silkworm, *Bombyx mori*: frequency, distribution, mutations, marker potential and their conservation in heterologous species. *Genetics* **169**: 197–214.
- Ramsay L, Maccaulay M, Cardle L, Morgante M, Degli Ivanisovich S, Maestri E et al. (1999). Intimate association of microsatellite repeats with retrotransposons and other dispersed repetitive elements in barley. *Plant J* **17**: 415–425.
- Reddy KD, Abraham EG, Nagaraju J (1999). Microsatellites in the silkworm, *Bombyx mori*: abundance, polymorphism, and strain characterization. *Genome* **42**: 1057–1065.
- Rychlik W (2000). *OLIGO: Primer Analysis Software*. Molecular Biology Insights. West Cascade, CO, USA.
- Temnykh S, DeClerck G, Lukashova A, Lipovich L, Cartinhour S, McCouch S (2001). Computational and experimental analysis of microsatellites in rice (*Oryza sativa* L.): frequency, length variation, transposon associations, and genetic marker potential. *Genome Res* **11**: 1441–1452.
- Tero N, Neumeier R, Gudavalli R, Schlötterer C (2006). *Silene tatarica* microsatellites are frequently located in repetitive DNA. *J Evol Biol* **19**: 1612–1619.
- Van't Hof AE, Zwaan BJ, Saccheri IJ, Daly D, Bot ANM, Brakefield PM (2005). Characterization of 28 microsatellite loci for the butterfly *Bicyclus anynana*. *Mol Ecol Notes* **5**: 169–172.
- Wilder J, Hollocher H (2001). Mobile elements and the genesis of microsatellites in dipterans. *Mol Biol Evol* **18**: 384–392.
- Williams BL, Brawn JD, Paige KN (2002). Highly polymorphic microsatellite loci for *Speyeria idalia* (Lepidoptera: Nymphalidae). *Mol Ecol Notes* **2**: 87–88.
- Wu Q, Andolfatto P, Haunerland NH (2001). Cloning and sequence of the gene encoding the muscle fatty acid binding protein from the desert locust, *Schistocerca gregaria*. *Insect Biochem Mol Biol* **31**: 553–562.
- Zhang DX (2004). Lepidopteran microsatellite DNA: redundant but promising. *TREE* **19**: 507–509.

Supplementary Information accompanies the paper on Heredity website (<http://www.nature.com/hdy>)