Genomic Gigantism: DNA Loss Is Slow in Mountain Grasshoppers

Douda Bensasson, Dmitri A. Petrov,† De-Xing Zhang,‡ Daniel L. Hartl,* and Godfrey M. Hewitt§*

*Department of Organismic and Evolutionary Biology, Harvard University; †Department of Biological Sciences, Stanford University; ‡Institute of Zoology, Chinese Academy of Sciences, Beijing, People's Republic of China; and §School of Biological Sciences, University of East Anglia, Norwich, England

Several studies have shown DNA loss to be inversely correlated with genome size in animals. These studies include a comparison between Drosophila and the cricket, Laupala, but there has been no assessment of DNA loss in insects with very large genomes. *Podisma pedestris,* the brown mountain grasshopper, has a genome over 100 times as large as that of Drosophila and 10 times as large as that of Laupala. We used 58 paralogous nuclear pseudogenes of mitochondrial origin to study the characteristics of insertion, deletion, and point substitution in *P. pedestris* and Italopodisma. In animals, these pseudogenes are ''dead on arrival''; they are abundant in many different eukaryotes, and their mitochondrial origin simplifies the identification of point substitutions accumulated in nuclear pseudogene lineages. There appears to be a mononucleotide repeat within the 643-bp pseudogene sequence studied that acts as a strong hot spot for insertions or deletions (indels). Because the data for other insect species did not contain such an unusual region, hot spots were excluded from species comparisons. The rate of DNA loss relative to point substitution appears to be considerably and significantly lower in the grasshoppers studied than in Drosophila or Laupala. This suggests that the inverse correlation between genome size and the rate of DNA loss can be extended to comparisons between insects with large or gigantic genomes (i.e., Laupala and Podisma). The low rate of DNA loss implies that in grasshoppers, the accumulation of point mutations is a more potent force for obscuring ancient pseudogenes than their loss through indel accumulation, whereas the reverse is true for Drosophila. The main factor contributing to the difference in the rates of DNA loss estimated for grasshoppers, crickets, and Drosophila appears to be deletion size. Large deletions are relatively rare in Podisma and Italopodisma.

Introduction

The brown mountain grasshopper (*Podisma pedestris*) has a genome that is seven times as large as that of humans and about the same size as that of the onion (calculated from the genome sizes given in Westerman, Barton, and Hewitt [1987] and Li [1997]). There is massive variation in the haploid genome sizes (C-values) of different eukaryotes, and this variation is not correlated with the amount of genic DNA these eukaryotes possess. The question of why this is so—known as the C-value paradox—has troubled biologists for over 50 years (Cavalier-Smith 1985). Several DNA-sequence studies suggest that genome size is correlated with the rate at which nucleotides are deleted in nongenic DNA (e.g., Graur, Shauli, and Li 1989; Petrov, Lozovskaya, and Hartl 1996; Ophir and Graur 1997; Petrov and Hartl 1998; Robertson 2000). Drosophila, with a genome size of 176 Mb, has a deletion rate that is much higher than those of the mammals studied (2,900–3,400 Mb) or Laupala crickets (1,910 Mb) (Petrov, Lozovskaya, and Hartl 1996; Petrov et al. 2000). In Drosophila, at least, rates of DNA deletion appear to be high enough to contribute significantly to the paucity of pseudogene sequences in the genome (Petrov, Lozovskaya, and Hartl 1996).

Most acridid grasshoppers have larger genomes (5,950–20,600 Mb) than other insects (98–8,900 Mb) and all mammals (1,420–5,680 Mb), and *P. pedestris* has one of the largest of these (18,150 Mb) (genome sizes approximated from Rees, Shaw, and Wilkinson [1978], Rasch [1983], Westerman, Barton, and Hewitt [1987], and Li [1997]). What are the characteristics of DNA deletion in an organism with such a large genome? Is the rate at which nucleotide deletions accumulate in Podisma pseudogenes significantly lower than those observed in Drosophila, crickets, or mammals? Does the correlation between genome size and nucleotide deletion rates extend to comparisons between species without streamlined genomes? We addressed these questions by characterizing nucleotide deletions in the sister genera Podisma and Italopodisma across 58 different pseudogene DNA sequences and comparing them with the published data for other species. The genome size of *Italopodisma* sp. is not known, but we assumed it to be similar to that of *P. pedestris.* Several of the Podisma and Italopodisma pseudogenes studied arose prior to the divergence of these recently diverged sister genera.

The pseudogenes used in this study are recently arisen nuclear copies of mitochondrial genes (hereinafter referred to as ''Numts'', following the abbreviation of Lopez et al. [1994]). Animal Numt sequences almost certainly lose their function when they escape the mitochondria (Gellissen and Michaelis 1987). They are ''dead on arrival.'' This makes them ideal for comparing rates of insertion and deletion (indel) mutation relative to rates of nuclear substitution mutation (Graur, Shauli, and Li 1989; Petrov, Lozovskaya, and Hartl 1996). Numt sequences evolve very differently from their paralogous mtDNA sequences, so nuclear mutations that arose since a Numt's loss of function can easily be dis-

Abbreviations: indel, insertion or deletion; Numt, nuclear mitochondrial DNA.

Key words: Acrididae, C-value paradox, Numt sequences, numtDNA, mtDNA, transposition, nuclear copies, horizontal gene transfer, dead-on-arrival (DOA) sequences.

Address for correspondence and reprints: Douda Bensasson, Hartl Lab, Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, Massachusetts 02138. E-mail: dbensass@oeb.harvard.edu.

Mol. Biol. Evol. 18(2):246–253. 2001
© 2001 by the Society for Molecular Biology and Evolution. ISSN: 0737-4038

tinguished from changes that arose in mitochondrial lineages. The Numts of Podisma and Italopodisma arose through many independent transfers to the nucleus (Bensasson, Zhang, and Hewitt 2000), so the nuclear changes they accumulate probably arose independently in each Numt. The many pseudogenes studied here are paralogous, so sequence composition varies in a similar way across each one, which allows the study of sequence composition effects on mutation. Numts are widespread in eukaryotes and abundant in mammals and grasshoppers (Zhang and Hewitt 1996), so this approach could be employed on other species.

Materials and Methods

Classification within the genus Italopodisma, based solely on subtle differences in the shape of the male genitalia, is known to be difficult (Harz 1975). For this reason the Italopodisma used in this study are referred to as *Italopodisma* sp. The Numt sequences used are paralogous to 643 bp of the mitochondrial ND5 region. Fifty-eight different ND5-like Numt DNA sequences (from two *Italopodisma* sp. and five *P. pedestris* individuals) and nine mtDNA sequences were employed in this study. GenBank accession numbers are as follows: mtDNA sequences for *P. pedestris* individuals, *Italopodisma* sp., *Parapodisma mikado*, and *Schistocerca gregaria*—AF085501–AF085507, AF085561; mtDNA sequence for *Locusta migratoria*—X80245; Numt sequences—AF085508–AF085538, AF085540–AF085545, AF085547–AF085550, AF085552–AF085560, AF085562–AF085565, AF085575–AF085578. All of the materials and methods used to generate these data are described in detail in Bensasson, Zhang, and Hewitt (2000) and are therefore outlined here only briefly.

Total genomic DNA was extracted from each individual. Oligonucleotide PCR primers were designed with DNA sequences matched to that of *L. migratoria* (GenBank accession number X80245; Flook, Rowell, and Gellissen 1995) in such a way that they would amplify a variable 643-bp portion of ND5 by annealing to stretches of sequence that are conserved among Orthopterans. These primers amplified paralogous ND5-like Numts, as well as mtDNA sequences, in all grasshoppers studied, although all PCR products appeared to be of the expected 688-bp size (643 bp plus primers). PCR products were cloned, and several (2–34) recombinant colonies were picked for each PCR product. The insert was amplified by PCR from each colony and sequenced from both strands. As the sequences used were not very divergent, they were aligned and checked by eye in MacClade, version 3.01 for Macintosh (Maddison and Maddison 1992).

A DNA extraction that enriches for mtDNA was successfully employed on three frozen individuals (the two Italopodisma and one of the *P. pedestris* individuals). When the PCR primers were employed on mitochondrial enriched DNA, ND5 products were almost exclusively of one type in each individual, and this was assumed to be the mtDNA sequence. However, the mitochondrial enrichment protocol was not successful for the other *P. pedestris* individuals studied, probably because they were ethanol-preserved and their mitochondria were therefore less well preserved (Dowling et al. 1996). For these other *P. pedestris* individuals, the mtDNA was assumed to be the type which contained no frameshift mutations and was present in high copy numbers.

Fifty of the 58 Numt sequences used were picked from PCR products that were amplified using a high-fidelity polymerase (*Pfu* polymerase). To confirm that the *Pfu* error was low enough for the purposes of this study, *Pfu* polymerase was used to amplify a PCR product from a colony with a known Numt insert instead of total genomic DNA. The product was cloned, 10 colonies were picked, and their inserts were sequenced from both strands. All 10 sequences were identical to the original insert sequence; therefore, misincorporations resulting from the techniques employed here were few. The remaining eight Numt sequences that were selected from *Taq*-amplified PCR products are labeled with a appended ‘‘t’’ (e.g., SpB5t). A total of two to three nucleotide substitutions and one indel would be expected to result from *Taq* errors in these eight sequences (estimated from the *Taq* error rates published in Kwiatowski et al. [1991]). Three or four *Taq* misincorporations would not affect the outcome of this study.

Separating Nuclear from Mitochondrial Lineage Changes

All insertions and deletions were assumed to occur in the noncoding (Numt) lineages for the following reasons: (1) 65 of the 66 indels observed in this study would have generated a frameshift mutation and thus would probably be highly deleterious in the mitochondrial lineage; (2) no indels had arisen in the mtDNA sequences since their ancestral divergence, although the sequences represent acridids of three different subfamilies; and (3) no indels arising from somatic mutation of mtDNA molecules could be detected. To study somatic mutation in mtDNA, PCR products were amplified, cloned, and sequenced from the mitochondrial enriched DNA of three individuals: 11 colonies with mtDNA inserts were sequenced from one individual, 5 from another, and 4 from the third. All mtDNA sequences within individuals were identical. If there are any somatic mutations in mtDNA molecules, their frequency is too low to affect this study.

Thirty-one of the Italopodisma Numts belong to a closely related family of Numts (Bensasson, Zhang, and Hewitt 2000). The nucleotide differences among these Numts are few. They differ from one another only by indels and one or two random point substitutions, but their majority-rule consensus sequence differs from the current mtDNA type by 13 point substitutions and no indels. It appears that the pseudogenes of this family are descended from the same mitochondrial sequence and that the differences among them arose in the nucleus. For a fuller discussion of this conclusion, see Bensasson, Zhang, and Hewitt (2000).

```
[                        10        20        30        40        50        60        70        80        90        100]
[                        .         .         .         .         .         .         .         .         .         .]

Fr Pp mt        TTATTTATGTGTGCGGGTTCAATAATTCATAATTTAAAGGATTCTCAGGATATTCGTTTTATAGGCTCACTTGTAAATTTTATACCTTTAACTTCAATTT
SpA Pp mt       ....................................................................................................
SpB Pp mt       ....................................................................................................
Fn Pp mt        ........A........G..................................................................................
SpC Pp mt       ........A........G..................................................................................
Itsp mt         ..G..C..A........A................................................T.TG...............................
Par m mt        ........A.....A...................................................T..GG....T....................T....
Sg mt           ........A.....A.....................G.............................A...A....T.....................G...
Lm mt           .....C........A..C..T..............GCGA.......................G..T...A..A.TC.......G...........T....
excluded sites    *  *  *    **  *  *    *         ****      *        *       *  * ***  * **       *     *        **
SpB4            ....................................................................................................
FrB1t           .................G..................................................................................
SpB1            .................G..................................................................................
SpB1t           .................G..................................................................................
SpC1t           .................G..................................................................................
SpC3t           .................G....................................C.............................................
FrA11           ........A........G..................................................................................
FrA2            ........A........G................................................T.................................
SpC2t           ........A.....A..G................................G.................................................
SpC11           ........A........G.......A.............................T.............................................
FrA1            ........A........G..................................................................................
FrC5            .......GA........GA.................................................................................
SpB2            ........A....T...G...........................................................?......................
SpB3t           ........A........A..................................................................................
SpB5t           ........A...T....A..................................................................................
FrA14           ........A....T...A.....................................G............................................
SpC3            ........A....T...A.....................................G............................................
FrA3            ........A........A.............................................G..T.................................
FrA18           ........A........A..G..........................................G..T.................................
FrB3t           .....C..A.....A..A................................................T.................................
FrA10           ........A.....A..A...........................T.................G..T....................C.............
ItA7            ..G..C..A........A................................................T.TG...............................
Itsp_fam        ..G..C...........A................................................T.TG...............................
ItA21           ........A.....A..A................................................T...A..........................G...
ItA43           ........AA...T...A................................................T...A..................A......T....
ItB1            ........A........A.......C...................T............A.......T...A........................GT....
SpC13           ..............CA................A.......T.......T.................T...A....T.....A..G.....G.....----.
SpC15           ..............A....................C.......................G.....T.T......T.......G.....G.....TG...
```

FIG. 1.—A small section of the full alignment to demonstrate the identification of ''unique'' (nuclear) mutations and shared (mitochondrial) point substitutions. Mitochondrial DNA sequences are aligned along the top, and are arranged in increasing order of divergence from the French *Podisma pedestris* mtDNA sequence (Fr *Pp* mt). Mitochondrial sequences were also included from three more distantly related grasshopper species whose Numts were not studied here: a *Parapodisma mikado* individual (*Par m* mt), *Schistocerca gregaria* (*Sg* mt), and *Locusta migratoria* (*Lm* mt). Numt sequences follow, also in increasing order of divergence from Fr *Pp* mt (not all the Numt sequences are shown here). Underlined sites were scored as point substitutions arising in the nucleus. * Nuclear mutations arising here would be shared with mtDNA differences or with more than one Numt sequence and consequently would not be identified; these sites are therefore excluded when counting the number of point substitutions. SpA *Pp* mt = mtDNA from *P. pedestris* Spanish individual A; SpB = *P. pedestris* Spanish individual B; Fn = *P. pedestris* Finnish individual; SpC = *P. pedestris* Spanish individual C; *It*sp mt = mtDNA from *Italopodisma* sp.; *It*A = Italopodisma individual A. Itsp_fam represents the majority consensus sequence of the *Italopodisma* sp. family of Numts.

Each of the remaining 27 Italopodisma and Podisma Numts diverged at different times from the Italopodisma and Podisma mitochondrial lineages (fig. 1). Some were 12% divergent from the current mtDNA sequences (at the base of the alignment in fig. 1), while others differed by as little as one mutation (at the top of the alignment). Numt sequences are thought to resemble ancestral mtDNA sequences (Fukuda et al. 1985; Hu and Thilly 1994), and a comparison of the sequences in figure 1 would support this. This is reflected in the tendency of the more divergent *Numt* pseudogenes in figure 1 to have a greater similarity to the more distantly related mtDNA sequences (*Par m*, *Sg*, and *Lm*) than to the Podisma and Italopodisma mtDNA sequences (Fr *Pp* through *It*sp).

Under mitochondrial evolution, sequences are functional and therefore under purifying selection; more changes are expected at silent sites (e.g., at third codon positions). Under nuclear evolution, Numt sequences have no function, and therefore the point substitutions they accumulate should be randomly distributed with respect to codon position. Most (if not all) of the 27 Numts arose from separate mitochondrial ancestors (Bensasson, Zhang, and Hewitt 2000). This was established by drawing pairwise comparisons between Numts. If their nucleotide differences showed significant codon position bias in their distribution (χ^2 test; $P < 0.05$), it was assumed that they had separate functional (mitochondrial) ancestors and were therefore transferred to the nucleus at separate times (Bensasson, Zhang, and Hewitt 2000).

It is clear from the small proportion of variable sites (fig. 1) that if nucleotide substitutions are distributed randomly (as expected under nuclear evolution), then multiple hits are unlikely. Nucleotide differences which are shared by more than one pseudogene usually represent differences acquired in a common ancestor. Since most Numts diverged from each other before their ancestor arrived in the nucleus (Bensasson, Zhang, and Hewitt 2000), shared differences probably arose in mitochondrial lineages (e.g., positions 3, 6, 9, 15, and 18 in fig. 1). For this reason, only unique substitutions (occurring at sites which were identical in all other nuclear or mitochondrial sequences) were counted as mutations occurring in the nucleus (fig. 1).

Of the 643 bp studied, there are 195 sites with variations in one of the mtDNA sequences or in two or more of the Numt sequences. Because only unique substitutions were counted as nuclear, any nuclear substitutions arising in these 195 sites were excluded. Therefore, the nuclear substitution frequency was estimated across what is effectively 448 bp.

This method for distinguishing nuclear and mitochondrial changes is similar in effect to a maximum-parsimony analysis. Maximum parsimony would assign all unique changes to the terminal branches of a tree, which would be occupied mostly by the dead-on-arrival lineages (Petrov, Lozovskaya, and Hartl 1996). However, shared substitutions resulting from multiple hits could also be ascribed to nuclear evolution, because parsimony makes no distinction between active (constrained) and inactive (pseudogene) lineages. A parsimony analysis of our data would ascribe some changes to nuclear mutations that we ascribe to mtDNA mutations, so parsimony would overestimate the number of nuclear mutations relative to the analysis presented here. Because we predicted a low number of deletions relative to point substitutions in grasshoppers, the most conservative test of this prediction would be to underestimate the number of point substitutions, since this would overestimate the rate of DNA loss. For the purposes of this study, our analysis was consequently more conservative than a maximum-parsimony analysis.

Testing the Separation of Nuclear and Mitochondrial Changes

Insect mitochondrial changes are expected to show a greater transition-to-transversion bias than changes accumulated in the nucleus (Tamura 1992; Moriyama and Powell 1996; Petrov and Hartl 1999). To test whether the separation of nuclear and mitochondrial changes was successful, we compared the transition-to-transversion bias observed for nuclear changes to that expected for mitochondria. For this comparison, we estimated mitochondrial transition-to-transversion bias by maximum likelihood for Podisma, Italopodisma, and closely related mitochondrial lineages (*L. migratoria* was not included in the analysis). The maximum-likelihood analysis of the mtDNA sequences of *P. pedestris* (five different sequences), Italopodisma, *P. mikado*, and *S. gregaria* was conducted with PAUP, version 4.0 (Swofford 1999). The settings used correspond to the Hasegawa, Kishino, and Yano (1985) model with rate heterogeneity. Empirical nucleotide frequencies were used and the gamma distribution was estimated using four rate categories. The estimated transition-to-transversion ratio was 1.7 (kappa = 4.8). This shows a strong bias in favor of transition mutations, as expected for insect mtDNA sequence (Tamura 1992).

Results

Effective Separation of Nuclear and Mitochondrial Nucleotide Substitutions

Examination of the 31 different Numts belonging to the Italopodisma family of Numts revealed 43 nucleotide differences at 40 different sites. These have been ascribed to changes arising under nuclear evolution. They appeared to be distributed randomly across the 643-bp region studied, as would be expected if these sequences were evolving under no selective constraints. Twelve of the substitutions arose at the first codon position, 19 at the second, and 12 at the third; this is not significantly different from 1:1:1 (χ^2 test; $P = 0.3$, df $= 2$).

A second, independent, test of whether these nucleotide substitutions have arisen in the nucleus was to examine the ratio of transitions to transversions. In insects, the observed bias in favor of transitions was much less under noncoding nuclear evolution than under mitochondrial evolution (Tamura 1992; Moriyama and Powell 1996; Petrov and Hartl 1999). There were 18 transitions and 25 transversions. This ratio was not significantly different from 1:2, which is the ratio expected when there is no bias in favor of transitions (χ^2 test; $P = 0.2$, df $= 1$). On the other hand, 18:25 was significantly different from the 1.7:1 estimated for mtDNA sequences in these grasshoppers (χ^2 test; $P = 0.004$, df $= 1$). This result confirmed that the changes scored as nuclear in this study were characteristic of the type of mutation expected in the nucleus.

Similarly, the 126 changes scored as nuclear in the remaining data (27 Numts) showed no significant bias in their positions within codons (χ^2 test; $P = 0.9$, df $= 2$). The numbers in the first, second, and third codon positions (47, 47, and 32, respectively) were very close to the numbers expected (45, 48, and 33). The expected numbers were based on a 1:1:1 ratio correcting for the fact that the numbers of first-, second-, and third-codon-position sites included in the analysis were 160, 172, and 116. In contrast, the sites of shared Numt differences that were excluded from the analysis (on the grounds that they occurred in mtDNA) were constrained by selection; the numbers for first, second, and third positions were 55, 42, and 98 (χ^2 test; $P < 1 \times 10^{-5}$, df $= 2$), suggesting that they were correctly assumed to be evolving under selective constraints in the mitochondria. The observed transition : transversion ratio among nuclear changes (45:81) also showed no significant bias (χ^2 test; $P = 0.6$, df $= 1$), but it was significantly different from the ratio estimated across mtDNA sequences only (χ^2 test; $P < 1 \times 10^{-9}$, df $= 1$). This result was also as expected for mutations arising in the nucleus.

Indel Hot Spots

Examination of Numt and mtDNA sequences revealed a 21-bp region (209–229 bp from the start of the GenBank sequences used) with an unusual base composition. This region was very A+T-rich and featured a stretch of T's up to 10 bp long even in functional mtDNA. Forty of the 56 deletions observed, and 3 of the 9 insertions, occurred in this region. Both of these rates are much higher than would be expected by chance (Fisher's exact test; $P = 3 \times 10^{-11}$ and $P = 1 \times 10^{-5}$, respectively). Most of the pseudogenes used in this study arose independently of each other (Bensasson, Zhang, and Hewitt 2000), so the high indel frequency observed in this region cannot be attributed to common ancestry. This region appears to be a major hot spot for indel mutations, but it does not have an elevated rate of point substitutions. Single-base repeats are well documented as indel hot spots (Levinson and Gutman 1987).

A second hot spot was also discovered 137–143 bp from the start of the GenBank sequences: of the 16 remaining deletions, four occurred within a stretch of six T's. This rate is significantly higher than would be expected by chance (Fisher's exact test; $P = 0.002$). Each of these four deletions arose in pseudogenes with separate mitochondrial origins (data not shown); therefore, they arose independently.

Close examination of Laupala and Drosophila sequences revealed no mononucleotide repeats of more than 5 bp for Drosophila and no regions with elevated indel frequencies. Laupala did have one unusual 9-bp region (three C's and six A's) at positions 131–139. This was also the only region showing a significantly elevated indel frequency (7 of 65 indels occurred here) (Fisher's exact test; $P = 0.0001$). There were therefore major differences among the taxa studied in the sizes and strengths of hot spots. These differences appear to have arisen as a result of the differences in the lengths and compositions of the original sequence studied. For species comparability, we excluded hot spots from our analysis, except where otherwise specified, but we show a separate analysis that includes them in table 1 to illustrate how they could affect results.

Estimating the Overall Rate of DNA Loss

Once a pseudogene loses its function, we expect it to accumulate mutations steadily with time, in which case the number of deletions, insertions, and nucleotide substitutions should be correlated. The data support this expectation: the numbers of nucleotide substitutions are correlated with the numbers of indels observed (excluding hot spots, Pearson's $r = 0.6$, df $= 56$, $P < 0.001$). However, if the Italopodisma family of pseudogenes all arose at approximately the same time, each pseudogene would have had the same period of time in which to accumulate mutations. Therefore, if the mechanisms by which they arise are independent, then we expect the number of point substitutions accumulated by each Numt to not be correlated with the number of indels observed. This expectation was met (Pearson's $r = -0.1$, df $= 29$, $P = 0.5$). Because we could not date the origin of the Numts studied, rates of DNA loss were measured relative to rates of point substitution.

When the Italopodisma family of Numts was excluded, the numbers of indels observed in grasshoppers were counted across 616 bp of each Numt, but the numbers of point substitutions were counted across 448 bp. To adjust for this discrepancy when calculating the ratios

Table 1
Comparison of Deletion and Insertion Profiles of Insects With Different Genome Sizes

	Podisma and Italopodisma[a]		Laupala		
	Including Hot Spots	Excluding Hot Spots	Including Hot Spot	Excluding Hot Spot	Drosophila[b]
Genome size (Mb)	18,150[c]	18,150[c]	1,910	1,910	176
Size of data set[d]	169 p, 56 d, 9 i	169 p, 12 d, 6 i	662 p, 48 d, 18 i	662 p, 45 d, 14 i	669 p, 87 d, 10 i
Mean ratio of deletions to point substitutions[e]	0.28 (0.19–0.4)	0.06 (0.03–0.11)	0.07 (0.05–0.1)	0.07 (0.05–0.09)	0.13 (0.11–0.18)
Mean ratio of insertions to point substitutions[e]	0.04 (0.02–0.09)	0.03 (0.01–0.1)	0.03 (0.02–0.04)	0.02 (0.01–0.05)	0.01 (0.01–0.03)
Mean deletion size (bp)[f]	1.5 (1.2–1.8)	1.6 (1.0–2.2)	6.7 (2.5–11)	7 (2.7–11)	24.9 (20–30)
Mean insertion size (bp)[f]	1.1 (0.9–1.3)	1.2 (0.9–1.5)	5.3 (1.7–8.9)	6.5 (2–11)	3.2 (1.8–4.6)
Mean rate of DNA loss (per bp substitution)[g]	0.38 bp (0.18–0.57 bp)	0.06 bp (−0.03–0.16 bp)	0.27 bp (−0.02–0.7 bp)	0.34 bp (−0.01–0.73 bp)	3.2 bp (2.1–4.4 bp)

[a] *Podisma pedestris* and *Italopodisma* sp.

[b] Data described in Petrov et al. (2000).

[c] The haploid genome size inferred from the C-value given for *P. pedestris* in Westerman, Barton, and Hewitt (1987).

[d] p = point substitutions; d = deletions; i = insertions.

[e] Ranges in parentheses are 95% confidence intervals (calculated by maximum-likelihood analysis).

[f] Ranges in parentheses are 95% confidence intervals of the sample means.

[g] Ranges in parentheses are 95% confidence intervals (calculated through 1,000 simulated replicates).

of insertions or deletions relative to point substitutions, the total number of deletions observed was corrected to 5.8 (i.e., (8/616) × 448), and the number of insertions was corrected to 2.2 (i.e., (3/616) × 448). Such an adjustment inflates the variance of our estimate and therefore is conservative in hypothesis testing. The 95% confidence interval for the ratio of deletions to point substitutions was based on a maximum-likelihood analysis of the total numbers of deletions and the total numbers of substitutions observed. This analysis assumed that probabilities of mutations (deletions, insertions, or point substitutions) followed Poisson distributions and that mutations accumulated linearly with time.

The mean rate of DNA loss per point substitution was calculated as DNA loss = (deletion size × deletion ratio) − (insertion size × insertion ratio) using the mean indel sizes and ratios given in table 1. To estimate the 95% confidence intervals for the rate of DNA loss, each of these four sample means was assumed to follow normal distributions, and the rates of DNA loss were calculated from means drawn randomly from these four distributions over 1,000 replicates.

The DNA Loss Observed for Grasshoppers Is Slower than That for Crickets or Drosophila

Podisma and Italopodisma appear to be losing nonessential DNA at a rate considerably slower than the rate of 0.34 bp per point substitution observed for Laupala crickets ($P < 0.001$) or 3.2 bp observed for Drosophila ($P < 0.001$) (table 1). This effect can be mostly attributed to differences in the sizes of deletions. Podisma and Italopodisma accumulate significantly smaller deletions than Laupala (Wilcoxon test; $P = 0.005$) or Drosophila (Wilcoxon test; $P < 0.001$). More specifically, differences in deletion size appear to result from a difference in the frequency distributions of deletion sizes (fig. 2). The proportion of deletions that are larger than 1 bp is significantly smaller in grasshoppers than in Laupala (G test; df = 1, $P = 0.005$) or Drosophila (G test; df = 1, $P < 0.001$).

Discussion

Indel hot spots can radically bias estimates of the overall frequency of indel events. The major hot spot observed accounts for twice as many indels as the rest of the Numt region studied. Care should be taken to identify indel hot spots, especially when the rates of DNA loss for species are estimated across different types of DNA sequence. The existence of hot spots is established more easily using Numts (or other independently generated paralogous sequences) to study mutation than through the study of unrelated DNA sequences. This is because Numts provide numerous independent natural tests of whether new mutations are randomly distributed throughout the sequence.

The extent of bias appears to depend on the length of the hot spot sequence. The major hot spot (21 bp of unusual sequence) accounted for many more indels (43 out of 65) than the second grasshopper hot spot, with 6 bp of unusual sequence (4 of the remaining deletions), or the 9-bp Laupala hot spot (7 indels out of 66). This implies that if hot spot regions are included, species differences in the original hot spot sequence could lead to artifactual differences in deletion frequency.

Although in grasshoppers the overall rate of deletion is not significantly higher than the insertion rate (in table 1, the mean rate of DNA loss is not significantly greater than zero), in the hot spot regions studied it is (when hot spot regions are included, the mean rate of DNA loss is significantly greater than zero; $P < 0.01$). Newly arisen pseudogenes will lose DNA from these indel mutation hot spots, so the degree to which they remain hot spots may fall. Therefore, indel accumulation in hot spots is probably not linear in time; these hot spots may be more common in ''young'' pseudogenes.

Differences in deletion size rather than frequency account for the largest component of the species differences in the estimated rates of DNA loss. The indel hot spots discovered did not contain deletions differing in size from those observed in the rest of the grasshopper data (Wilcoxon test; $P = 0.9$). Even if the many indels that arose in the hot spots were included in our analysis, the estimated rate of DNA loss in grasshoppers would still be considerably (and significantly) lower than that observed in Drosophila, although not significantly different from that in crickets. The hot spot observed in the Laupala data is too mild to significantly affect estimates of DNA loss (table 1).

All other reported comparisons have shown that species with smaller genomes have higher rates of DNA loss (e.g., Graur, Shuali, and Li 1989; Petrov, Lozovskaya, and Hartl 1996; Ophir and Graur 1997; Petrov and Hartl 1998; Petrov et al. 2000). With hot spots excluded, our data suggest that Laupala crickets (which have 10-fold less DNA than Podisma) lose nonessential DNA faster than *P. pedestris* grasshoppers. The low rate of DNA loss in Podisma is consistent with the reported pattern of DNA loss and suggests that the pattern may also apply to comparisons between species whose genomes are not as streamlined as that of Drosophila. The estimated rate of DNA loss in grasshoppers (less than 0.19 bp lost per nucleotide substitution) suggests that a site is less likely to be deleted than it is to change through a point substitution. In grasshoppers, therefore, the accumulation of point mutations is a more potent force for obscuring ancient pseudogenes than the accumulation of indels, while the reverse is true for Drosophila.

Petrov et al. (2000) showed that the frequencies of small deletions are similar between Laupala and Drosophila. *Caenorhabditis elegans,* with its small genome (86 Mb), shows a high proportion of large deletions, similar to those observed in Drosophila (176 Mb) and in contrast to those observed in humans (3,400 Mb) (Robertson 2000). The data presented here also suggest that it is the sizes and frequencies of large (>1 bp) deletions that vary among species with different genome sizes (table 1 and fig. 2). In grasshoppers, no deletions were observed that were larger than 4 bp. Perhaps deletions are generated by more than one mechanism: one mechanism is relatively constant between species and

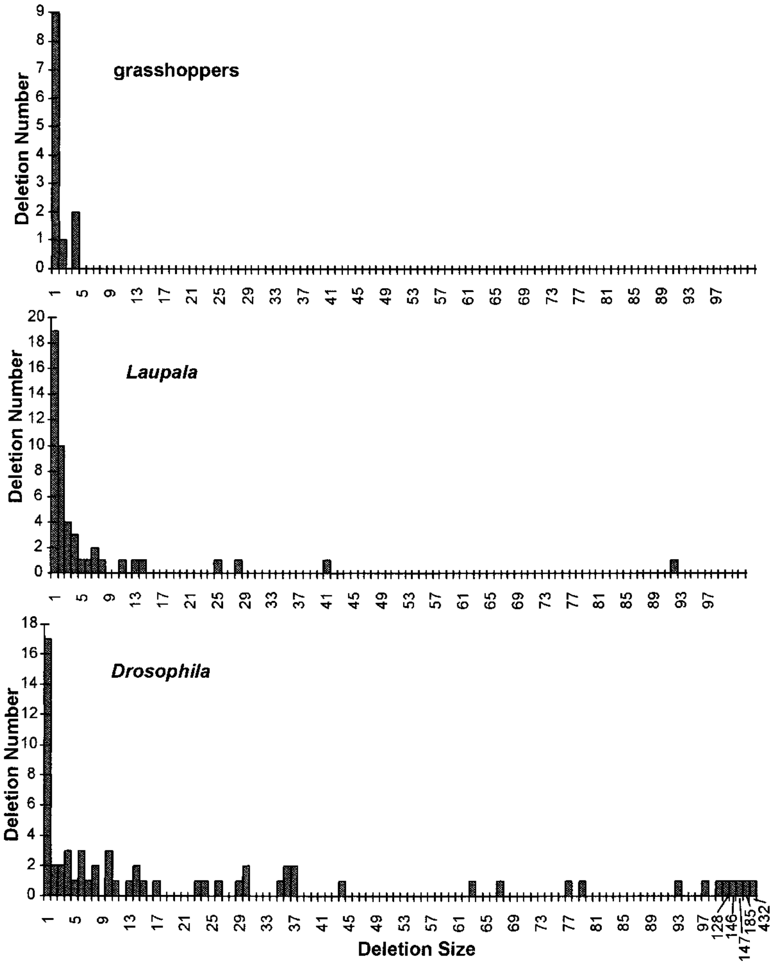

FIG. 2.—The frequency distribution of deletion sizes in the grasshoppers Podisma and Italopodisma compared with data from Laupala and *Drosophila melanogaster* (Petrov et al. 2000).

generates predominantly small (single-nucleotide) deletions (e.g., as expected through slipped-strand mutation), while a second mechanism generates a broader range of deletion sizes (e.g., as expected through unequal sister chromatid exchange or unequal crossing over) and is more active in organisms with smaller genomes.

The 58 Numts analyzed here represent at least 40 kb of inserted DNA; in the time since they arose, they have sustained a net loss of only about 10 bp (0.025%) through indel accumulation. This rate may seem slow, but if applied across the colossal grasshopper genome, well over 4.6 Mb of DNA loss would be expected since the Numts used here arose ([169 point substitutions/ 40,000 bp Numt DNA] × 18,150,000,000 bp genome size × 0.06 bp mean DNA loss).

This and previous studies (e.g., Graur, Shuali, and Li 1989; Petrov, Lozovskaya, and Hartl 1996; Ophir and Graur 1997; Petrov and Hartl 1998; Petrov et al. 2000) have demonstrated that relative rates of DNA loss differ across species. But what about absolute amounts of DNA loss? Surprisingly, it seems that absolute amounts of DNA loss are comparable in the species studied here. If grasshoppers have 100 times as much nonessential DNA as Drosophila and 10 times as much as Laupala (not unreasonable considering their differences in genome size), then indels will accumulate across 100 times as many sites as in Drosophila and across 10 times as many as in Laupala. For example, when each species has accumulated a million point substitutions across its genome, Drosophila will have lost 3.2 Mb of DNA, Laupala will have lost 3.4 Mb (0.34 × 10), and Podisma will have lost 6 Mb (0.06 × 100).

It is interesting that the significantly different rates of deletion per nucleotide site (brought about by different patterns of mutation) have arisen in different species of insects, while the absolute rate of DNA loss per genome remains on the same order of magnitude. Perhaps the absolute amounts of DNA gain due to large insertions (transposable elements, Numts, etc.) in these species are also similar, which would suggest that their genome sizes are near or at equilibrium. More information on the rate of DNA acquisition, and on other mechanisms of DNA loss, would be needed to evaluate this possibility.

Even if the genome sizes of the insects discussed here were presently at equilibrium, the observed differences in deletion rates may have played an important role in determining the present genome sizes. If the absolute rate of DNA gain remained constant, and if (for the sake of argument) there was no selection acting on genome size, a change in the relative rate of indel accumulation would drive a change in genome size until the same total amount of DNA was lost across the genome once more. For example, if grasshoppers were to acquire a 40-fold greater rate of DNA loss (the same relative rate of indel accumulation as Drosophila) in the absence of selection, and if the amount of DNA gain were to remain unchanged, the amount of nonessential DNA in grasshopper genomes would shrink. As genome size decreases, the absolute amount of DNA lost also decreases until it reaches its original value at an approximately 40-fold smaller genome size. In light of the extremely slow rate of DNA loss actually observed, it is not surprising that the genomes of Podisma and Italopodisma are gigantic.

Acknowledgments

This work was partly funded by an NERC studentship, the University of East Anglia, and a Royal Society-Fulbright postdoctoral fellowship (to D.B.) and a grant (GM58423) from the National Institutes of Health (to D.L.H.). We thank Jeff Townsend, Carlos Bustamante, and Yun Tao for helpful discussions, and two anonymous reviewers for thoughtful comments on an earlier version of the manuscript.

LITERATURE CITED

BENSASSON, D., D.-X. ZHANG, and G. M. HEWITT. 2000. Frequent assimilation of mitochondrial DNA by grasshopper nuclear genomes. Mol. Biol. Evol. **17**:406–415.

CAVALIER-SMITH, T. 1985. The evolution of genome size. John Wiley, New York.

DOWLING, T. E., C. MORITZ, J. D. PALMER, and L. H. RIESEBERG. 1996. Nucleic acids III: analysis of fragments and restriction sites. Pp. 249–320 *in* D. M. HILLIS and C. MORITZ, eds. Molecular systematics. Sinauer, Sunderland, Mass.

FLOOK, P. K., C. H. F. ROWELL, and G. GELLISSEN. 1995. The sequence, organization and evolution of the *Locusta migratoria* mitochondrial genome. J. Mol. Evol. **41**:928–941.

FUKUDA, M., S. WAKASUGI, T. TSUZUKI, H. NOMIYAMA, K. SHIMADA, and T. MIYATA. 1985. Mitochondrial DNA-like sequences in the human nuclear genome. J. Mol. Biol. **186**: 257–266.

GELLISSEN, G., and G. MICHAELIS. 1987. Gene transfer: mitochondria to nucleus. Ann. N.Y. Acad. Sci. **503**:391–401.

GRAUR, D., Y. SHUALI, and W.-H. LI. 1989. Deletions in processed pseudogenes accumulate faster in rodents than in humans. J. Mol. Evol. **28**:279–285.

HARZ, K. 1975. The Orthoptera of Europe II. Dr. W. Junk, the Hague, the Netherlands.

HASEGAWA, M., H. KISHINO, and T. YANO. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J. Mol. Evol. **32**:443–445.

HU, G., and W. G. THILLY. 1994. Evolutionary trail of the mitochondrial genome as based on human 16S rDNA pseudogenes. Gene **147**:197–204.

KWIATOWSKI, J., D. SKARECKY, S. HERNANDEZ, D. PHAM, F. QUIJAS, and F. J. AYALA. 1991. High fidelity of the polymerase chain reaction. Mol. Biol. Evol. **8**:884–887.

LEVINSON, G., and G. A. GUTMAN. 1987. Slipped-strand mispairing: a major mechanism for DNA sequence evolution. Mol. Biol. Evol. **4**:203–221.

LI, W.-H. 1997. Molecular evolution. Sinauer, Sunderland, Mass.

LOPEZ, J. V., N. YUHKI, R. MASUDA, W. MODI, and S. J. O. O'BRIEN. 1994. *Numt*, a recent transfer and tandem amplification of mitochondrial DNA to the nuclear genome of the domestic cat. J. Mol. Evol. **39**:174–190.

MADDISON, W. P., and D. R. MADDISON. 1992. MacClade: analysis of phylogeny and character evolution. Version 3.0. Sinauer, Sunderland, Mass.

MORIYAMA, E. N., and J. R. POWELL. 1996. Intraspecific nuclear DNA variation in *Drosophila*. Mol. Biol. Evol. **13**: 261–277.

OPHIR, R., and D. GRAUR. 1997. Patterns and rates of indel evolution in processed pseudogenes from humans and murids. Gene **205**:191–202.

PETROV, D. A., and D. L. HARTL. 1998. High rate of DNA loss in the *Drosophila melanogaster* and *Drosophila virilis* species groups. Mol. Biol. Evol. **15**:293–302.

———. 1999. Patterns of nucleotide substitution in *Drosophila* and mammalian genomes. Proc. Natl. Acad. Sci. USA **96**: 1475–1479.

PETROV, D. A., E. R. LOZOVSKAYA, and D. L. HARTL. 1996. High intrinsic rate of DNA loss in *Drosophila*. Nature **384**: 346–349.

PETROV, D. A., T. A. SANGSTER, J. S. JOHNSTON, D. L. HARTL, and K. L. SHAW. 2000. Evidence for DNA loss as a determinant of genome size. Science **287**:1060–1062.

RASCH, E. M. 1983. DNA ''standards'' and the range of accurate DNA estimates by Feulgen absorption microspectrophotometry. Pp. 137–166 *in* R. R. COWDEN and F. W. HARRISON, eds. Advances in microscopy. A. R. Liss, New York.

REES, H., D. D. SHAW, and P. WILKINSON. 1978. Nuclear DNA variation among acridid grasshoppers. Proc. R. Soc. Lond. B Biol. Sci. **202**:517–525.

ROBERTSON, H. M. 2000. The large *srh* family of chemoreceptor genes in *Caenorhabditis* nematodes reveals processes of genome evolution involving large duplications and deletions and intron gains and losses. Genome Res. **10**:192–203.

SWOFFORD, D. L. 1999. PAUP*: phylogenetic analysis using parsimony (*and other methods). Version 4.0. Sinauer, Sunderland, Mass.

TAMURA, K. 1992. The rate and pattern of nucleotide substitution in *Drosophila* mitochondrial DNA. Mol. Biol. Evol. **9**:814–825.

WESTERMAN, M., N. H. BARTON, and G. M. HEWITT. 1987. Differences in DNA content between two chromosomal races of the grasshopper *Podisma pedestris*. Heredity **58**:221–228.

ZHANG, D.-X., and G. M. HEWITT. 1996. Nuclear integrations: challenges for mitochondrial DNA markers. Trends Ecol. Evol. **11**:247–251.

ELIZABETH KELLOGG, reviewing editor

Accepted October 26, 2000