JOURNAL OF EXPERIMENTAL ZOOLOGY (MOL DEV EVOL) 306B:539–550 (2006)

Multiple Copies of Coding as Well as Pseudogene *c-mos* Sequence Exist in Three Lacertid Species

MIHAELA PAVLICEV* AND WERNER MAYER
Laboratory of Molecular Systematics, Natural History Museum Vienna, 1010 Vienna, Austria

ABSTRACT The analysis of a 581 bp section of the nuclear gene *c-mos* revealed multiple copies of putative functional sequences as well as pseudogenes in three closely related lacertid species *Lacerta laevis*, *L. kulzeri* and *L. cyanisparsa*. A phylogenetic analysis of *c-mos* in comparison with a molecular phylogeny based on the mitochondrial cytochrome *b* gene supports our findings. The study also provides new insights into the phylogenetic relationships of *L. cyanisparsa* and *L. laevis*.

Pseudogenes of the three species share 11 single-nucleotide substitutions, a 1 bp deletion and a premature stop codon but differ by group-specific mutations. This result suggests that the *c-mos* gene has become duplicated and subsequently silenced already in the common ancestor of the three species. Sequence divergence suggests that the duplication and the loss of function occurred in the late Miocene/early Pliocene, i.e., about 5 million years ago. Indications of gene conversion are discussed.

We suggest that future studies using *c-mos* for phylogenetic studies should provide evidence for the orthology of the sequences compared. *J. Exp. Zool. (Mol. Dev. Evol.) 306B:539–550, 2006.*

Gene duplication has long been considered an important mechanism in supplying raw material for evolutionary change (Ohno, '70). In many organisms duplicated genes have been found in high proportions; in eukaryotes, estimates range from 30% to 65% of the total gene number (Zhang, 2003). Following gene duplication, the copies can either retain the same function (genetic redundancy), or one of the copies can change or lose function (i.e., silencing). Unless extra gene product is advantageous, true genetic redundancy is generally considered evolutionarily unstable. Therefore, different explanations for the maintenance of paralogous genes have been proposed (e.g., Nowak et al., '97; Force and Lynch, '99). Duplicate gene pairs are observed sometimes to evolve asymmetrically (e.g., Conant and A. Wagner, 2003; G.P. Wagner et al., 2005). The existence of equivalent duplicates can also be seen as a potentially (but not necessarily) transitional state to one of the alternatives: adoption of a partial (subfunctionalization), related or new function (cooption or neofunctionalization). All possibilities feature divergent evolution of paralogs (e.g., globin and opsin gene families, Hox clusters; see Ohta, '93, '94; Holland, '99; Lynch and Force, 2000; Briscoe, 2001; G.P. Wagner et al., 2003; review in Zhang, 2003). Finally, in the case of silencing, the time from the duplication event to the loss of function can be substantial (e.g., Bailey et al., '78), particularly if one considers partial (gradual) loss of independently mutable subfunctions (Lynch and Force, 2000). In general, as long as it is transcribed, a gene might be still constrained because of potentially deleterious protein products resulting from amino acid replacement (Hughes, '94; Conant and A. Wagner, 2003 and references therein). The copies not being transcribed anymore (pseudogenes) are freed completely from selective pressure and thus evolve at a faster rate compared to the original functional sequence. All mutations in a pseudogene are considered neutral, including introduction of internal stop codons, modification of splice signals and insertions/deletions. The proportion of non-synonymous substitutions thus increases, approaching random probability (approximately 75% due to the properties of the genetic code).

*Correspondence to: M. Pavlicev, Natural History Museum Vienna, Laboratory of Molecular Systematics, Burgring 7, 1010 Vienna, Austria. E-mail: mihaela.pavlicev@nhm-wien.ac.at
Received 13 February 2006; Accepted 14 March 2006
Published online 31 May 2006 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/jez.b.21110.

The presence of multiple gene copies implies the risk of non-orthologous comparisons in the phylogenetic studies. However, the pseudogenes, when identified and analyzed, may also provide additional information for phylogenetic inference.

The presence of equivalent multiple functional copies cannot be ruled out for any gene, unless the complete genome sequence is available. In most cases, multiple copies of the marker gene are identified in the course of phylogenetic studies if the additional copies are highly divergent or degenerated (i.e., pseudogenes). This was the case in a study, where we used the nuclear *c-mos* gene as a phylogenetic marker.

The *c-mos* proto-oncogene is an intronless nuclear gene that codes for a serine/threonine kinase involved in the regulation of the cell cycle in vertebrates (Gebauer and Richter, '97; Sagata, '97). Because of its slow evolutionary rate, Graybeal ('94) and Lovette and Bermingham (2000) suggested *c-mos* as a potential marker for investigating phylogenetic relationships at higher taxonomic levels. It has since been widely used for phylogenetic research, either alone or in multilocus comparisons. Most of the existing data sets concern birds (e.g., Cooper and Penny, '97; Butorina and Solovenchuk, 2004; Overton and Rhoads, 2004; Voelker and Spellman, 2004) and reptiles (e.g., Saint et al., '98; Harris et al., '99, 2001; Brehm et al., 2001; Carranza et al., 2002, 2004; Townsend et al., 2004).

In this paper, we present evidence for the existence of several functional and non-functional copies of the *c-mos* gene in the genomes of the three closely related lacertid species *Lacerta laevis*, *L. kulzeri* and *L. cyanisparsa*. Furthermore, we characterize the pseudogene sequences and propose a scenario about their origin and evolution.

METHODS

DNA extraction, PCR amplification and sequencing

A 581 bp section of the *c-mos* gene was analyzed. Total genomic DNA was extracted from deep frozen or alcohol-preserved tissues following a standard phenol–chloroform procedure (Sambrook et al., '89). Samples are listed in Table 1.

Amplifications of all PCR fragments were performed in 25 μl reaction mixtures containing PCR buffer with 1.5 mM $MgCl_2$, 0.2 mM of each dNTP, 0.4 μM of each PCR primer, 0.5 units of *Taq* polymerase (Amersham, UK). Reaction conditions comprised an initial denaturation step of 2 min at 94°C, 35 cycles of 10 sec at 95°C, 15 sec at annealing temperature, 50 sec at 72°C and a final extension step of 7 min at 72°C. Negative and positive PCR controls were included in all PCR amplifications. Sequences and annealing temperatures of primers used in the study are given in Table 2.

We performed direct sequencing of PCR products as well as sequencing of the cloned PCR fragments. The advantage of the first method is that it can reveal the presence of diverged copies by the appearance of fixed heterozygosity in the population. A disadvantage is, however, that the overlapping graphical representations of single copies cannot be disentangled; therefore, cloning is necessary.

In the case of direct sequencing, we repeated PCR amplifications using the preamplified *c-mos* segment as a template. The nested primers used for reamplification were L1zmos and Hcmos1. We further used CMS-77L and CMS-482H as sequencing primers for direct sequencing.

In the case of cloning, gel-purified (QIAquick® Gel extraction kit, Qiagen, Netherlands) PCR products were cloned using the TA vector (TOPO TA Cloning Kit, Invitrogen, Carlsbad, CA).

For all individuals included in the *c-mos* tree, we analyzed the total sequence of the mitochondrial cytochrome *b* (*cyt b*) gene (1,143 bp) to illustrate phylogenetic relationships at mitochondrial (mt) level. Sequencing was performed by MWG (Ebersberg, Germany). We repeated the procedures with newly amplified PCR products of several randomly chosen samples to double-check the results.

Sequence analysis

The sequences were aligned and edited manually with the software BioEdit (Version 7.0.1; Hall '99). All *c-mos* sequences were searched manually for obvious gene conversions between functional and the most divergent non-functional sequences. Converted sequences were excluded, as far as recognized (Table 3). In the *c-mos* gene tree only those individuals were included for which both functional and non-functional sequences were identified. We applied Bayesian inference (MrBayes, version 3.1; Huelsenbeck and Ronquist, 2001) to construct phylogenetic trees, using the general time reversible model. For both the *c-mos* and the *cyt b* gene tree, two million generations were run with a sampling frequency of 100. Of the 20,000 sampled trees, the majority consensus tree was built over the final 5,000 trees. Gaps were treated as missing characters by the algorithms used for the inference of phylogenetic relationships.

TABLE 1. List of samples analyzed for the c-mos gene and their geographical origin

Sample code	Species	Geographical origin	FG	N-FG
GZ-1[1]	*Lacerta cyanisparsa*	Gaziantep, S Turkey	1	1
GZ-3	*Lacerta cyanisparsa*	Gaziantep, S Turkey	0	2
GZ-4[1]	*Lacerta cyanisparsa*	Gaziantep, S Turkey	1	3
RH-3[1]	*Lacerta cyanisparsa*	Al Barah, Syria	1	2
FP-9[1]	*Lacerta danfordi*	Camliyayla, Icel, S Turkey	1	0
JD-2	*Lacerta kulzeri*	Jebel Druz, S Syria	1	0
JD-7	*Lacerta kulzeri*	Jebel Druz, S Syria	1	0
KC-3	*Lacerta kulzeri*	Antilibanon, Bludan, SW Syria	0	1
KP-1	*Lacerta kulzeri*	Petra, W Jordan	0	1
LU-5[1]	*Lacerta kulzeri*	Maalula, Antilibanon, SW Syria	1	1
LU-6[1]	*Lacerta kulzeri*	Maalula, Antilibanon, SW Syria	1	1
UY-2	*Lacerta kulzeri*	Jebel Sanin, Uyum as Sanine, Lebanon	0	2
UY-3	*Lacerta kulzeri*	Jebel Sanin, Uyum as Sanine, Lebanon	1	0
ZB-3	*Lacerta kulzeri*	Jebel Barouk, Lebanon	0	3
1DJ-2	*Lacerta laevis*	Jebel Barouk, Lebanon	1	1
BZ-2	*Lacerta laevis*	Camliyayla, Icel, S Turkey	3	1
CK-1[1]	*Lacerta laevis*	Ansari Mountains, NW Syria	1	1
DM-2	*Lacerta laevis*	Damascus, Syria	0	1
HA-2[1]	*Lacerta laevis*	Harbiye, Hatay, S Turkey	2	1
JR-1[1]	*Lacerta laevis*	Zubiya, NW Jordan	1	1
KC-1	*Lacerta laevis*	Bloudan, Antilibanon, SW Syria	0	1
LH-1	*Lacerta laevis*	Samandagi, Hatay, S Turkey	3	4
LR-1	*Lacerta laevis*	Andirin, Kahramanmaraş, S Turkey	0	2
LR-2[1]	*Lacerta laevis*	Andirin, Kahramanmaraş, S Turkey	1	1
LV-5[1]	*Lacerta laevis*	Polis, W Cyprus	3	2
LY-2[1]	*Lacerta laevis*	Yayladagi, S Turkey	1	1
LH-3	*Lacerta laevis*	Reyhanli, Hatay, S Turkey	1	0
LH-5	*Lacerta laevis*	Reyhanli, Hatay, S Turkey	1	0
LH-7	*Lacerta laevis*	Reyhanli, Hatay, S Turkey	0	1
NN-1[1]	*Lacerta laevis*	Tannourine, Lebanon	2	1
JE-2[1]	*Lacerta laevis*	Jerusalem, Israel	2	1
JE-3	*Lacerta laevis*	Jerusalem, Israel	1	0
JN-1[1]	*Lacerta laevis*	Junie, Lebanon	1	1
BB-3[1]	*Lacerta laevis*	Baalbek, Lebanon	3	1
BY-2[1]	*Lacerta laevis*	Byblos, Lebanon	2	2
PZ-1[1]	*Lacerta laevis*	Yarpuz, Osmanye, S Turkey	1	2
BH-1[1]	*Lacerta laevis*	Bcharrèe, Lebanon	1	1
BH-2[1]	*Lacerta laevis*	Bcharrèe, Lebanon	3	1
BH-3	*Lacerta laevis*	Bcharrèe, Lebanon	0	2
BH-5[1]	*Lacerta laevis*	Bcharrèe, Lebanon	1	5

All DNA samples are stored at the Natural History Museum Vienna (NHMW), Austria. FG: number of analyzed functional sequences; N-FG: number of analyzed non-functional sequences.
[1]Samples used in the gene trees.

TABLE 2. Primer sequences and the corresponding annealing temperatures

Name	Sequence	T	Purpose
Hcmos3	5′-GGT GAT GGC AAA TGA GTA GAT-3′	55°C	Initial PCR
L1zmos	5′-CTA GCT TGG TGT TCT ATA GAC TGG-3′	55°C	Initial PCR, reamplification PCR
Hcmos1	5′-GCA AAT GAG TAG ATG TCT GCC-3′	56°C	Reamplification PCR
CMS-77L	5′-CTA CGT ACC ATG GAG CTA C-3′	56°C	Sequencing
CMS-482H	5′-TTG GGA ACA TCC AAA GTC TC-3′	56°C	Sequencing
LKcm-dR	5′-GAT GCC AAA CGG TTC TTA CTG C-3′	57°C	Localization
LKcm-uF	5′-GTG GTC CTG AAC TCC TTA AAG G-3′	57°C	Localization

TABLE 3. Frequency distribution of sequence types among the total number of sequences

	Functional	Non-functional	Obvious conversion	Total
Lacerta laevis N	15	10	9	34
Lacerta laevis S	20	25	22	67
Lacerta cyanisparsa	3	8	13	24
Lacerta kulzeri	5	9	6	20

The sequences were classified as either functional, non-functional or converted sequences.

GenBank accession numbers: *cyt b*: from DQ461744 to DQ461765, inclusive; coding *c-mos*: from DQ461713 to DQ461743, inclusive; pseudogene *c-mos*: from DQ461682 to DQ461712, inclusive.

RESULTS

Identification of multiple c-mos sequences

In the course of a phylogenetic study (Mayer and Pavlicev, 2005; Mayer and Pavlicev, unpublished), where we analyzed samples from 31 lacertid genera by direct sequencing of PCR products, a high degree of polymorphism was found in the southern population of *L. laevis*. This polymorphism was first detected by multiple bands in the electropherogram and further analyzed by sequencing of cloned PCR products for the present study. Altogether three species, *L. laevis*, *L. kulzeri* and *L. cyanisparsa*, were found to possess multiple copies of *c-mos* and the analysis was extended to additional individuals. Besides putative functional copies, in one population non-functional copies were recognized by a premature stop codon and two deletions of 1 and 7 bp, respectively (Fig. 3). Subsequently, screening of clones revealed both types of the *c-mos* gene in the remaining populations of the three species, whereby the non-functional copies were characterized by different population-specific deletions. Altogether, a total of 101 clones from 28 individuals of *L. laevis*, 20 clones from nine individuals of *L. kulzeri* and 24 clones of four individuals of *L. cyanisparsa* were sequenced.

For reasons that will become clear below, we analyze northern and southern populations of *L. laevis* separately because they appear not to form a monophyletic group. We thus refer to groups, rather than species.

Distribution of gene copies

To investigate the relative localization of the multiple copies to each other (e.g., clusters or tandem repeats), we designed outward directed primers (LKcm-dR and LKcm-uF; Table 2) which should bind in putatively neighboring pseudogenes and enable amplification of the sequence between them. This strategy has the advantage of being fairly simple, but it is limited by the maximum length of the amplifiable segment. The functionality of the primers used for this test has been previously confirmed on the same samples by combining each with the corresponding primer originally used for amplifying *c-mos* sequence. This test produced partial *c-mos* sequences.

The test for the spatial distribution of duplicated copies within the genome of *L. laevis* revealed that the distances between paralogous sequences probably exceed 3.5 kb, which we consider a very conservative estimate of the maximum amplifiable sequence length in the PCR. This finding suggests that different copies actually occupy distant locations within the genome.

Non-functional c-mos sequences (pseudogenes)

We concentrate in the following on the obviously non-functional *c-mos* sequences, identified by the presence of premature stop codons and deletions that cause a frameshift. The comparison between the functional and non-functional sequences revealed that the pseudogenes of each of the four groups have specific characteristics. They all share a 1 bp deletion, a premature stop codon and 11 single-nucleotide substitutions. Additionally, group-specific substitutions and deletions are found. The differences with respect to the functional genes are summarized in Table 4a.

Tables 4b and c present information on the type of substitutions, the ratios of transitions to transversions and the ratios of non-synonymous vs. synonymous substitution rates. The functional and non-functional sequences were compared within groups. To do this, we calculated the total number of synonymous (N_S) and non-synonymous (N_N) sites for the functional sequences and for the pseudogenes (Table 4b). Applying the reading frame of the functional gene, we classified the observed substitutions in the pseudogenes as either synonymous (M_S) or non-synonymous (M_N) as shown in Table 4c. The rate of synonymous substitutions (dS) is calculated as observed synonymous substitutions (M_S) divided by the total synonymous sites (N_S). Correspondingly, the rate of the non-synonymous substitutions (dN) is revealed by dividing the observed non-

synonymous substitutions (M_N) by non-synonymous sites (N_N). Because of the low divergence between the sequences, the values were not corrected for multiple substitutions at a single site. The ratio ω between these two rates (d*N*/d*S*; Table 4c) is interpreted as a measure of selective pressure. In the case of a neutrally evolving sequence, the ratio between these two rates is expected to approach 1. High values of ω (i.e., a higher rate of non-synonymous substitution) reflect directional selection, whereas low values of ω suggest stabilizing selection in the sequence compared.

In the analyzed pseudogene, ω is close to 1 in all groups except in *L. kulzeri*. If statistically significant, the result for *L. kulzeri* would suggest directional selection on the putative pseudogene.

To estimate the probability of observed ratio of non-synonymous to synonymous substitutions (M_N/M_S; Table 4c) in the pseudogenes of each group, we calculated the group-specific a priori probabilities of non-synonymous (P_N) and synonymous (P_S) substitutions per site from the functional sequence for each particular group (Table 4b). We used these a priori probabilities to estimate the probabilities of observed ratio P (M_N/M_S) in each group by binomial distribution (Table 4c). In all cases, this probability exceeds 0.05, meaning that the hypothesis of neutral evolution of the presumptive pseudogenes cannot be rejected in any of the groups, including *L. kulzeri*.

TABLE 4A. Characterization of the pseudogenes

	N	Ni	Nt	Deletion length	Stop codon	M_N+M_S
L. laevis N	10	7	9	1	1	16
L. laevis S	25	16	16	1, 7	1	17
L. cyanisparsa[1]	8	4	5	1, 22, (2), (38)	1	17
L. kulzeri	9	6	7	1,12, 3, 6	1	23

N is the total number of pseudogenes obtained across all clones; N_i is the number of individuals examined; N_t is the number of different variants of the pseudogenes; M_N+M_S is the total number of group-specific substitutions (non-synonymous and synonymous).
[1]Two types of pseudogene were found, one with two additional deletions.

TABLE 4B. Comparison of the pseudogenes to coding c-mos sequences

	Coding sequence			Pseudogene		
	N_N	N_s	P_N	N_N	N_s	P_N
L. laevis N	446.7	132.3	0.772	441.3	131.7	0.770
L. laevis S	446.7	132.3	0.772	435.7	128.3	0.772
L. cyanisparsa	446.7	132.3	0.772	424.0	125.0	0.772
L. cyanisparsa[1]				(391.3)	(115.7)	(0.838)
L. kulzeri	446.7	132.3	0.772	430.0	128.0	0.771

N_N is total number of non-synonymous sites per sequence; N_s is total number of synonymous sites per sequence. The proportion of non-synonymous sites [$P_N = N_N/(N_N+N_s)$] reflects the overall probability of non-synonymous mutations. Correspondingly, the probability of synonymous mutation for this stretch is $P_S = (1-P_N)$.
[1]The statistics is calculated separately for the second distinct type of the pseudogene found in *L. cyanisparsa*.

Functional c-mos sequences

In Table 5, the polymorphic sites among the functional sequences within each group are examined in more detail. For simplicity, only a single sequence per individual is included (intra-individual variation, see below); therefore, the number of sequences per group (*N*) does not correspond to the total number of functional sequences present in the data set, but rather to the total number of individuals for which functional sequences are available. All sites that proved

TABLE 4C. Characterization of the pseudogene types by the group

	N	M_N+M_S	*T*s/*T*v	M_N/M_S	d*N*	d*S*	ω	$P(M_N/M_S)$
L. laevis N	10	16	10/6	13/3	0.029	0.023	1.261	0.228
L. laevis S	25	17	11/6	13/4	0.030	0.031	0.968	0.222
L. cyanisparsa	8	17	12/5	13/4	0.031	0.032	0.969	0.222
L. cyanisparsa[1]					(0.033)	(0.035)	(0.943)	(0.165)
L. kulzeri	9	23	12/11	20[2]/2	0.047	0.016	2.875	0.067

N is the number of pseudogenes analyzed. M_N and M_S are the observed numbers of non-synonymous and synonymous changes, respectively, common to all pseudogenes within a group. d*N* and d*S* are the non-synonymous and synonymous substitution rates, respectively, ω is the ratio d*N*/d*S*. In a neutrally evolving sequence ω is expected to approach 1. In general, values significantly >1 suggest directional selection, values <1 stabilizing selection. $P(M_N/M_S)$ is the probability of the observed ratio of non-synonymous to synonymous substitutions, given the species-specific probabilities of non-synonymous (and synonymous) sites (Table 4a), under assumption of neutrality.
[1]The statistics is calculated separately for the second distinct type of the pseudogene found in *L. cyanisparsa*.
[2]Single codon is affected by two substitutions.

TABLE 5. Characterization of the intraspecific variation in coding sequences calculated for each group

	N	M_N+M_S	T_s/T_v	M_N/M_S	dN	dS	ω	$P(M_N/M_S)$
L. laevis N	8	10	8/2	6/4	0.013	0.030	0.433	0.017
L. laevis S	13	8	8/0	6/2	0.013	0.015	0.867	0.016
L. cyanisparsa[1]	3	0	0	0	—	—	—	—
L. kulzeri	5	7	4/3	$6^2/1$	0.013	0.007	1.857	0.004

N is the number of coding *c-mos* sequences compared; M_N and M_S are the observed numbers of synonymous and non-synonymous substitutions, respectively; T_s/T_v is the transition/transversion ratio; M_N/M_S the ratio between synonymous and non-synonymous substitutions. dS and dN are the synonymous and non-synonymous substitution rates, respectively; ω is non-synonymous to synonymous substitution rate ratio (dN/dS). $P(M_N/M_S)$ is the probability of the observed ratio of non-synonymous to synonymous substitutions, given the species-specific probabilities of non-synonymous (and synonymous) sites (Table 4a), under assumption of neutrality.

TABLE 6. Ranges of intra-individual sequence divergence in %

	Compared sequence pair		
	Functional vs. functional (%)	Functional vs. non-functional (%)	Non-functional vs. non-functional (%)
Lacerta laevis N	0.0–0.9	2.8–3.9	0.0–0.5
Lacerta laevis S	0.0–0.5	3.0–4.3	0.0–1.1
Lacerta cyanisparsa	No data	2.9–3.3	0.0–0.4
Lacerta kulzeri	No data	4.8–5.0	0.0–0.07

variable in the particular group are counted as polymorphic. Apparently, converted sequences were excluded from the data set and will be addressed later in the text.

The ratio ω in the functional copies of *c-mos* is also shown in Table 5. Note that this ratio is somewhat higher in *L. kulzeri*.

The probability of observing the particular ratios of non-synonymous to synonymous substitutions (P (M_N/M_S); Table 5) under neutrality is <0.05 in all groups. This means that the neutrality of the sequences can be rejected for all groups, supporting their classification as coding. Nevertheless, it is important to note that the variation in each group revealed at this level can be accounted for by both allelic polymorphism within the group as well as multiple copies of the *c-mos* gene within single organisms. Therefore, in the next section we examine the variation of copies within single individuals.

Within-individual variation in non-functional and functional sequences

Our findings support occurrence of multiple copies of putatively functional as well as non-functional *c-mos* sequences in all three species. Therefore, it is interesting to see how much these paralogous sequences differ within single individuals. In Table 6 we present the ranges of within-individual variation in sequence divergence. The values are descriptive only, since they are based on a small sample of individuals.

At most, three distinct copies of a functional sequence were found within a single individual, and up to five non-functional copies (Table 1). Single-nucleotide polymorphisms theoretically can originate by amplification errors during the PCR. However, due to the high frequency of observed polymorphisms as well as their repeatability (see discussion), we exclude this option as an explanation.

The most obvious differences were found in *L. cyanisparsa*, where two distinct pseudogene types differing by two deletions are found even in the same individual. In this particular case the polymorphism could not be explained by conversion.

Owing to the sample size, within-individual polymorphisms among the presumed **functional** paralogs could be studied only in *L. laevis* (seven individuals). The number of polymorphic sites across the clones within a single individual ranged from 1 to 6 per 581 bp *c-mos* stretch (non-synonymous/synonymous ratios 3/1, 5/1, 2/3, 1/0, 1/0, 1/5 and 0/1). As in the previous examination of the within-group variation, part of the variation may represent allelic variation in heterozygous individuals.

Converted c-mos sequences

The comparative analysis of the sequences revealed numerous cases of apparent gene conversion. Converted sequences were defined as recombinant sequences consisting partly of known functional and non-functional sequences occurring in the respective group sample. The observed numbers are listed in Table 3. All recognized converted sequences were excluded from further analysis. In principle, the occurrence of recombined sequences can be explained also by artificial PCR recombination (i.e., jumping PCR; Meyerhans et al., '90; Judo et al., '98). Nevertheless, the high numbers of converted sequences imply natural gene conversion. We examine the sources of recombination more thoroughly in the Discussion section.

Phylogenetic relationships among populations

From the total data set, 21 individuals representing all three species were selected to evaluate phylogenetic relationships. We used only those individuals for whom both, a functional *c-mos* gene and a *c-mos* pseudogene were available. Figure 1 shows the relationships among the *c-mos* sequences. It can be seen that there is a well-supported clade of pseudogenes of the three species, which is separated from the clade of functional sequences. The general topologies within these two main clades are similar. Yet this topology does not correspond to the current taxonomic status of the species concerned. According to both sequences, the individuals of *L. laevis* form two distinct geographic groups, designated as *L. laevis* North (N) and *L. laevis* South (S) in Figure 1. *L. cyanisparsa* clusters with some individuals of *L. laevis* N. This means that *L. laevis* (and even its northern subgroup alone) is paraphyletic in the tree. Because of this clear distinction, we treated the two populations *L. laevis* N and *L. laevis* S as separate groups in our study. We also treated *L. cyanisparsa* as a separate group due to the specific characteristics of its pseudogenes, despite its position within the *L. laevis* N cluster of the gene tree (a detailed phylogeographic study of the *laevis–cyanisparsa* complex is in preparation; Mayer et al., unpublished). In *L. kulzeri*, the coding sequence fails to provide sufficient statistical resolution for its placement, whereas the pseudogene sequence indicates its close relationship to *L. laevis* S.

For this study, we have classified all sequences with intact reading frames as ''functional'', although their functionality was not tested experimentally. However, the topology and the branch lengths of the corresponding clades in the *c-mos* gene tree support this classification.

To test the inference of genetic relationships among the studied species as revealed by the *c-mos* sequences, we analyzed yet another marker gene, the mt *cyt b* (Fig. 2) from the same individuals that have been used above for the *c-mos* sequence phylogeny (Fig. 1). With respect to the group relationship, the topology of the *cyt b* tree is in accordance with the functional *c-mos* tree. The separation of the northern and southern *L. laevis* populations is confirmed as well as the clustering of several individuals of *L. laevis* N with *L. cyanisparsa*. As in the functional *c-mos* branch of the *c-mos* gene tree, the position of *L. kulzeri* in the *cyt b* tree is not resolved. Thus in spite of the evidence of multiple *c-mos* copies the sequence phylogeny based on the functional copies of *c-mos* is still in accordance with the phylogeny based on *cyt b*.

The concordance between the topologies of the two gene trees further supports the interpretation of deviated sequences as pseudogenes. A comprehensive lacertid phylogeny based on the functional *c-mos* sequences and the *recombination activating gene* (*rag-1*) is presented elsewhere (Mayer and Pavlicev, 2005; Mayer and Pavlicev, unpublished).

DISCUSSION

In several individuals of *L. laevis*, *L. kulzeri* and *L. cyanisparsa,* two coexisting types of *c-mos* sequences were discovered, one functional and one with impaired function due to deletions, nonsynonymous nucleotide substitutions and a premature stop codon. This can be considered sufficient evidence for the existence of a pseudogene. However, the question remains, how many intact copies of *c-mos* are really present in the studied genomes, apart from the obvious non-functional copies? Furthermore, how many different non-functional copies are there?

Although this study does not provide a direct answer to these questions, we make indirect inferences from the observed variability between paralogs.

Within-individual polymorphism

Repeated cloning of the *c-mos* gene from single individuals revealed multiple copies of both the obvious non-functional and the putative functional

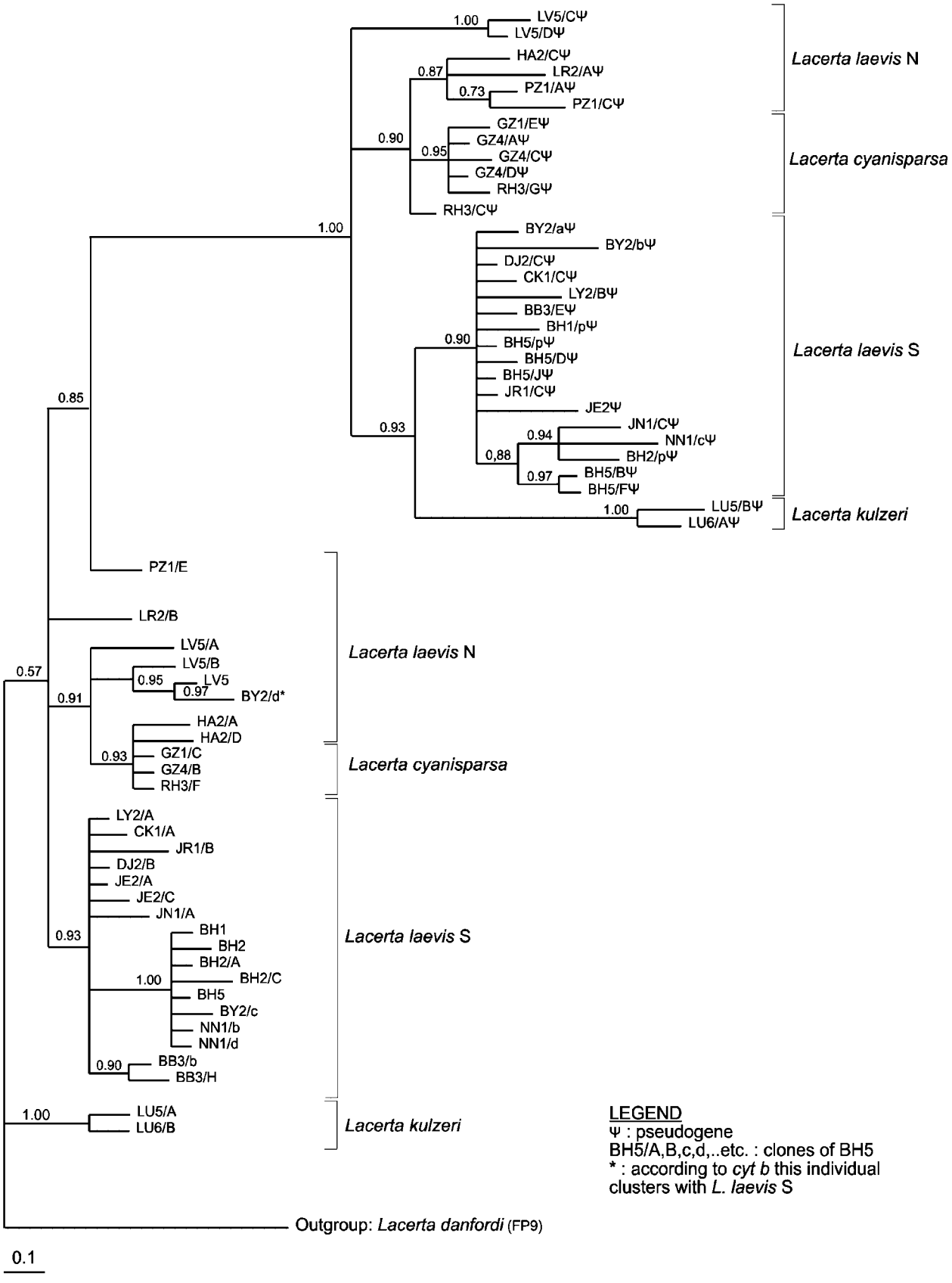

Fig. 1. Gene tree based on *c-mos* sequences.

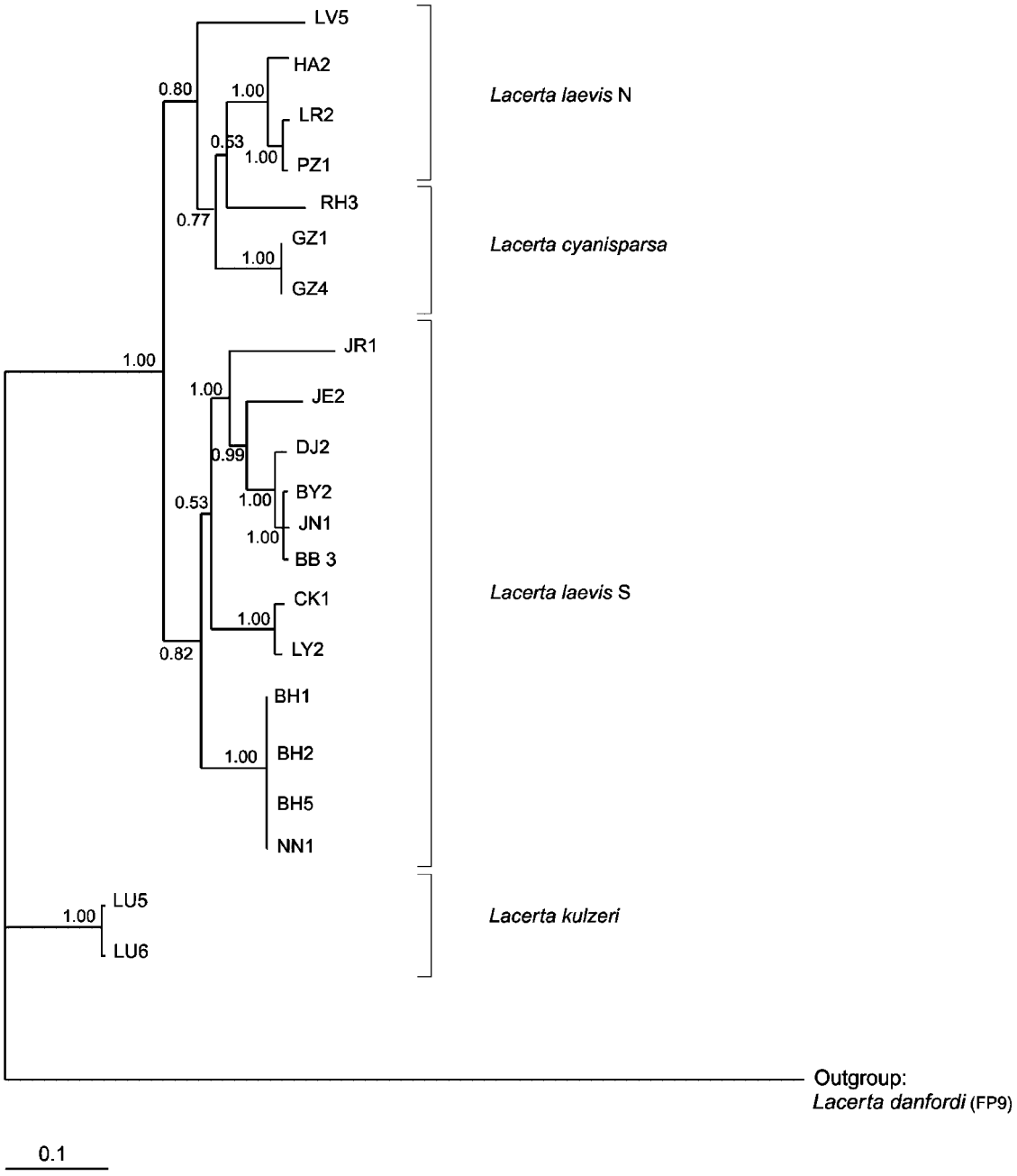

Fig. 2. Gene tree based on *cyt b* sequences.

sequences, differing in single-nucleotide substitutions. Theoretically, at least six ways to attain such a polymorphic situation are conceivable: real genetic polymorphism due to (i) polyploidy, (ii) heterozygosity, (iii) gene duplication (with or without recombination caused by gene conversion); and experimental artefacts due to (iv) jumping PCR, (v) contamination or (vi) random mutations generated during the PCR. In the following, we discuss each possibility with the exception of polyploidy, which can be ruled out for these taxa (In den Bosch et al., 2003) and will not be discussed here further (but see, e.g., Evans et al., 2005), and contamination, which we consider unlikely to have produced the observed pattern.

For a single-copy gene, the maximum number of alleles in the nuclear genome of a diploid organism is two and the difference between these two alleles can be calculated from the number of heterozygous nucleotide positions. However, at each single

```
                1        10        20        30        40        50        60        70        80        90       100*
CODING          GATCAGTTGTGCCTGCTGCACCCCCTAGGCTCTGGTGGCTTTGGTTCTGTCTACAAGGCTACATACCATGGAGCTACAGTGGCTGTAAAGCAGGTAAAAAA**
L. laevis S     GATCAGTTGTGCCTGCTGCACCCCCTAGGTTCTGGTGGCTATGGTTCTGTCTACAAGGCTACATACTATGGAGCTACAGTGGCTATAAAGCAGGTAA-AAA**
L. laevis N     GATCAGTTGTGCCTGCTGCACCCCCTAGGTTCTGCTGGCCATGGTTCTGTCTACAAGGCTATATACCATGGAGCTACAGTGGCTATAAAGCAGGTAA-AAA**
L. cyanisparsa A GATCAG----------------------GTTCTGGTGGCTATGGTTCTGTCTACAAGGCTATATACCATGGAGCTACAGTGGCTATAAAGCAGGTAA-AAA**
L. cyanisparsa B GATCAG----------------------GTTCTGGTGGCTATGGTTCTGTCTACAAGGCTATATACCATGGAGCTACAGTGGCTATAAAGCAGGTAA-AAA**
L. kulzeri      GATCAGTTGTGCCTGCTGCACCCCCTAGGCTCTGGTGGCTGTGGTTCTGTCTAC-----------TATGGAGCTACAGTGGCTATAAAGCAGGTAA-AAA**

*   180       190       200       210       220 * 340       350       360       370       380  ** *530       540
*TGTGGTGCGTATAGTAGCTGCTAGCACATGTGCCCCTGGCAATC** *GTGTCTGGGCTACTCCTGTGACATTATGGCCGGCTTAGTGTTTCTC** *CATACACCCACCGTGCTCCT**
*TGTGGTGTGTATACTAGCTGCTAGCACATGT-------GCAATC** *GTGTCTGAGCTACTCCTGTGACATTATGGCAGGCTTAGTGTTTCTC** *CATACACCCACTGTGCTCCT**
*TGTGGTGTGTATACTAGCTGCTAGCACATGTGCCCCTGGCAATC** *GTGTCTGAGCTACTCCTGTGACATTATGGCAGGCTTAGTGTTTCTC** *CATACACCCACCGTGCTCCT**
*TGTGGTGTGTATACTAGCTGCTAGCACATGTGCCCCTGGCAATC** *GTGT-----------------------------------TCTC** *CATACACCCACCGTGCTCCT**
*TGTGGTGTGTATACTAGCTGCTAGCACATGTGCCCCTGGCAATC** *GTGTCTGAGCTACTCCTGTGACATTATGGCAGGCTTAGTGTTTCTC** *CATACACCCACCGTGCTCCT**
*TGTG---TGTATACTAGCTGCTAGCACATGTGCCCCTGGCAATC** *GTGTCTGAGCTACTCCTGTGACATTATGGCAGGCTTAATTTTTCTC** *CATAC------TGTGGTCCT**
```

Fig. 3. Alignment of the representative pseudogene sequences of each group against the coding sequence to demonstrate the pattern of deletions. Note two different types of pseudogenes in both *L. laevis* and *L cyanisparsa*. Asterisks denote the parts where the alignment has been shortened.

site, only two different nucleotides are possible, one for each allele. Thus one would expect to find at most two different haplotypes of a gene within a single individual, each with a certain combination of the character states over all polymorphic sites. This was not the case and thus can be taken as an indication of multiple gene copies (either with or without conversion) or jumping PCR.

The products of recombination events such as jumping PCR and gene conversion are detected as additional haplotypes. Considering four possible character states per polymorphic site and n polymorphic sites, the recombination by jumping PCR or gene conversion can generate 4^n different haplotypes with respect to these sites. For gene conversion to occur, multiple copies of a gene must exist. It is very improbable that all converted sequences can be recognized. Nevertheless, given that the functional genes are clearly diverged from the pseudogenes, at least recombinant sequences between these two extremes can be identified. Indications of gene conversion between pseudogene and functional sequences occurred in all three species (Table 3). Further cases are probable, especially since the converted sections are not always easily ascribed to a particular sequence type, and since conversion can occur between pseudogenes and even already converted sequences, as well. Undetected gene conversions (either with functional or non-functional copies) may blur the relationships when reconstructing phylogenetic trees.

Jumping PCR (Meyerhans et al., '90) produces in vitro recombination of sequences and can explain the generation of artificial ''haplotypes''. Among other factors, the frequency of jumping PCR presumably depends on the quality of the DNA used in the PCR, fragmented DNA being especially prone to this phenomenon. The use of freshly preserved material in this study, allowing amplification of rather long fragments, makes this scenario unlikely. Besides, the empirically estimated rate for the occurrence of PCR recombination is quite low (for *Taq* polymerase 1% in 12 doublings per 282 bp [Judo et al., '98]) and may have accounted only marginally for the number of copies found in the present study. Similarly, random errors of DNA polymerases are known to be responsible for single substitutions during PCR (*quasi* mutations; error rates of up to 8×10^{-6} for *Taq* polymerase have been documented [Cline et al., '96]), but the detected number of polymorphisms exceeds the proposed rates. Furthermore, our results were confirmed by repeated PCR experiments.

Pseudogenes

When compared to the functional genes, pseudogenes are expected to demonstrate equal rates of non-synonymous and synonymous substitutions (ω approaches 1), which is an indication of neutral evolution. In this study, we could unambiguously demonstrate neutrality for all groups examined. The deviation from the expected distribution was highest in *L. kulzeri* where the rate of non-synonymous substitutions exceeds that of synonymous substitutions three-fold. Such a result could indicate that directional selection on the defective *c-mos* copy acts only in this species, conflicting with the inference of pseudogene status from the presence of a premature codon in all species as a synapomorphic character. As has been mentioned, the ratio between non-synonymous and synonymous substitution rates is higher also among the functional *c-mos* paralogues of *L. kulzeri*. Nevertheless, all observed ratios of non-synonymous to synonymous substitutions

(M_N/M_S) in pseudogenes, including the one in *L. kulzeri*, fall within 95% ($P>0.05$) probability limits under the assumption of neutrality for the pseudogene; thus, the neutrality of these sequences cannot be rejected.

It should be noted, however, that our method of detecting pseudogenes has three important shortcomings: (i) the mutations that accumulate between duplication and silencing are not neutral, as long as the gene is still transcribed. This may affect the M_N/M_S ratio (the magnitude of this effect being correlated with the length of time between duplication and silencing); (ii) the uncertainty whether the presumably functional sequence to which the pseudogene is being compared is in fact the functional gene (especially if partial sequences are studied as in the present case); and (iii) the uncertainty of detection and exclusion of all converted sequences. In this study, the classification of the sequences as either pseudogenes or functional sequences has been supported by the sequence phylogeny itself and its comparison with the phylogeny derived from another marker sequence.

From the distribution of the pseudogenes in the groups studied and from the phylogenetic relationships among these groups, we conclude that the initial duplication of the *c-mos* (or its segment) originated before the common ancestor of the three species *L. laevis*, *L. kulzeri* and *L. cyanisparsa*. It may have acquired a single-nucleotide deletion leading to a frameshift and a change of an amino acid codon to a stop codon, mutations observed in all three species (Fig. 3). Therefore, the additional copy seems to have been silenced before the groups diverged. After that the pseudogenes evolved independently, accumulating species-specific deletions and substitutions. In the case of *L. laevis*, the 7 bp deletion has occurred either after the split between the northern and southern populations or has existed before and has become fixed only in the southern part of the species range. Similarly, *L. cyanisparsa* was most likely isolated from the rest of *L. laevis* N prior to the origin of its specific deletions, since these deletions were not found in any of the clones from *L. laevis* N. Multiple copies of the pseudogene presumably arose from subsequent duplication events of the pseudogene, after the segregation of the groups concerned. The low divergence of pseudogene copies within individuals and populations suggests recent duplications or repeated gene conversions. The observation of high numbers of obviously chimeric sequences (Table 3) between pseudogenes and functional sequences supports the assumption of a high rate of conversion.

Given that the *c-mos* duplication is absent in all other closely related species, we assume that both the duplication and the silencing occurred in the common ancestor of *L. laevis*, *L. kulzeri* and *L. cyanisparsa*. We therefore place both events in upper Miocene or early Pliocene (according to the dating based on mt sequences; Mayer and Pavlicev, unpublished).

Within-group polymorphism

Given the existence of both multiple functional and non-functional *c-mos* copies, intraspecific polymorphism of orthologous copies is hard to address since the existence of paralogs makes orthologous comparisons uncertain. Therefore, and due to a sample size that does not allow conclusions at the population level, it is not evident from this study what portion of the within-group polymorphism may be explained by allelic variation in the populations.

The ratios between non-synonymous and synonymous substitution rates (ω) in functional genes (Table 5) are in all groups lower than the corresponding ratios in pseudogenes (Table 4c). This supports the inference of selection against non-synonymous substitutions in putative functional sequences.

Conclusion

To conclude, the presented study of pseudogenes among *c-mos* gene sequences addresses the risks the undetected multiple copies of a gene impose for a phylogenetic inference. Even more important, the study demonstrates the potential information, the copies can contribute to the phylogenetic study, if detected and analyzed.

ACKNOWLEDGMENTS

We thank Elisabeth Haring and Wilhelm Pinsker for numerous discussions throughout the study, Günter Wagner for a critical view at the earlier version of the manuscript and the anonymous reviewer for constructive suggestions that greatly improved the original text. We also thank Josef Schmidtler and Wolfgang Bischoff for providing the samples.

LITERATURE CITED

Bailey GS, Poulter RTM, Stockwell PA. 1978. Gene duplication in tetraploid fish: model for gene silencing at unlinked duplicated loci. PNAS 75:5575–5579.

Bosch HAJ in den, Odierna G, Aprea G, Barucca M, Canapa A, Capriglione T, Olmo E. 2003. Karyological and genetic variation in Middle Eastern lacertid lizards, *Lacerta laevis* and the *Lacerta kulzeri* complex: a case of chromosomal allopatric speciation. Chromosome Res 11:165–178.

Brehm A, Jesus J, Pinheiro M, Harris DJ. 2001. Relationships of scincid lizards (*Mabuya* spp; Reptilia: Scincidae) from the Cape Verde islands based on mitochondrial and nuclear DNA sequences. Mol Phylogenet Evol 19:311–316.

Briscoe AD. 2001. Functional diversification of lepidopteran opsins following gene duplication. Mol Biol Evol 18: 2270–2279.

Butorina OT, Solovenchuk LL. 2004. The use of *c-mos* nuclear gene as a phylogenetic marker in Tetraonidae birds. Russ J Genet 40:1080–1084.

Carranza S, Arnold EN, Mateo JA, Geniez P. 2002. Relationships and evolution of the North African geckos, *Geckonia* and *Tarentola* (Reptilia: Gekkonidae), based on mitochondrial and nuclear DNA sequences. Mol Phylogenet Evol 23:244–256.

Carranza S, Arnold EN, Amat F. 2004. DNA phylogeny of *Lacerta* (*Iberolacerta*) and other lacertine lizards (Reptilia: Lacertidae): did competition cause long-term mountain restriction? Syst Biodivers 2:57–77.

Cline J, Braman JC, Hogrefe HH. 1996. PCR fidelity of *Pfu* DNA polymerase and other thermostable DNA polymerases. Nucleic Acids Res 24:3546–3551.

Conant GA, Wagner A. 2003. Asymmetric sequence divergence of duplicate genes. Genome Res 13:2052–2058.

Cooper A, Penny D. 1997. Mass survival of birds across the Cretaceous-Tertiary boundary: molecular evidence. Science 275:1109–1113.

Evans BJ, Kelley DB, Melnick DJ, Cannatella DC. 2005. Evolution of *RAG-1* in polyploid clawed frogs. Mol Biol Evol 22:1193–1207.

Force A, Lynch M. 1999. Preservation of duplicate genes by complementary, degenerative mutations. Genetics 151: 1531–1545.

Gebauer F, Richter JD. 1997. Synthesis and function of *Mos*: the control switch of vertebrate oocyte meiosis. Bioessays 19:23–28.

Graybeal A. 1994. Evaluating the phylogenetic utility of genes: a search for genes informative about deep divergences among vertebrates. Syst Biol 43:174–193.

Hall T. 1999. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/ NT. Nucleic Acids Symp Ser 41:95–98.

Harris DJ, Sinclair EA, Mercader NL, Marshall JC, Crandall KA. 1999. Squamate relationships based on C-*mos* nuclear DNA sequence. Herpetol J 9:147–151.

Harris DJ, Marshall JC, Crandall KA. 2001. Squamate relationships based on C-*mos* nuclear DNA sequences: increased taxon sampling improves bootstrap support. Amphibia-Reptilia 22:235–242.

Holland PWH. 1999. Gene duplication: past, present and future. Semin Cell Dev Biol 10:541–547.

Huelsenbeck JP, Ronquist F. 2001. Mr. Bayes: Bayesian inference of phylogenetic trees. Bioinformatics 17:754–755.

Hughes AL. 1994. The evolution of functionally novel proteins after gene duplication. Proc Biol Sci 256:119–124.

Judo MSB, Wedel AB, Wilson C. 1998. Stimulation and suppression of PCR-mediated recombination. Nucleic Acids Res 26:1819–1825.

Lynch M, Force A. 2000. The probability of duplicate gene preservation by subfunctionalization. Genetics 154:459–473.

Lovette I, Bermingham E. 2000. *C-mos* variation in songbirds: molecular evolution, phylogenetic implications, and comparisons with mitochondrial differentiation. Mol Biol Evol 17: 1569–1577.

Mayer W, Pavlicev M. 2005. Nuclear DNA sequences confirm the basal phylogeny of the family Lacertidae proposed by Harris, Arnold and Thomas (1998). Abstract, 13th Meeting of the European Society for Herpetology, Bonn, September 27–October 2, 2005.

Meyerhans A, Vartanian J-P, Wain-Hobson S. 1990. DNA recombination during PCR. Nucleic Acids Res 18: 1687–1691.

Nowak MA, Boerlijst MC, Cooke J, Smith JM. 1997. Evolution of genetic redundancy. Nature 388:167–171.

Ohno S. 1970. Evolution by gene duplication. Berlin: Springer-Verlag.

Ohta T. 1993. Pattern of nucleotide substitution in growth hormone-prolactin gene family: a paradigm for evolution by gene duplication. Genetics 134:1271–1276.

Ohta T. 1994. Further examples of evolution by gene duplication revealed through DNA sequence comparisons. Genetics 138:1331–1337.

Overton LC, Rhoads DD. 2004. Molecular phylogenetic relationships based on mitochondrial and nuclear gene sequences for the Todies (Todus, Todidae) of the Caribbean. Mol Phylogenet Evol 32:524–538.

Sagata N. 1997. What does *Mos* do in oocytes and somatic cells? Bioessays 19:13–21.

Saint KM, Austin CC, Donnellan SC, Hutchinson MN. 1998. *C-mos*, a nuclear marker useful for squamate phylogenetic analysis. Mol Phylogenet Evol 10:259–263.

Sambrook J, Fritsch EF, Maniatis T. 1989. Molecular cloning. A laboratory manual, 2nd edition. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press.

Townsend T, Larson A, Louis E, Macey JR. 2004. Molecular phylogenetics of squamata: the position of snakes, amphisbaenians, and dibamids, and the root of the squamate tree. Syst Biol 53:735–757.

Voelker G, Spellman GM. 2004. Nuclear and mitochondrial DNA evidence of polyphyly in the avian superfamily Muscicapoidea. Mol Phylogenet Evol 30:386–394.

Wagner GP, Amemiya C, Ruddle F. 2003. Hox cluster duplications and the opportunity for evolutionary novelties. PNAS 100:14603–14606.

Wagner GP, Takahashi K, Lynch V, Prohaska SJ, Fried C, Stadler PF, Amemiya C. 2005. Molecular evolution of duplicated ray finned fish HoxA clusters: increased synonymous substitution rate and asymmetrical co-divergence of coding and non-coding sequences. J Mol Evol 60:665–676.

Zhang J. 2003. Evolution by gene duplication: an update. TREE 18:292–298.