

J. W. Arntzen

Classification and molecular organization of satellites elucidated by phylogenetic network analysis – examples from *Triturus* salamanders and *Palorus* beetles

Received: 8 February 2001 / Accepted: 6 June 2002 / Published online: 16 August 2002
© Springer-Verlag 2002

Abstract A phylogenetic network of 244 satellite DNA sequences across five species of aquatic salamanders (genus *Triturus*) revealed four types of satellite DNAs in a ‘p’-shaped 1-2*-3-4-2* arrangement. Analysis of dimer and trimer DNA sequences revealed a prevalence of homosequential (e.g. 1-1, 2-2) and particular (1-4 and 2-3) heterosequential repeat motifs. Genetic diversity across types and species phylogeny indicated that type 1 and type 4 are derived from types 2 and 3. Support was also found for alternating motifs in *Palorus* flour beetle tandem repeats. The results were statistically significant, whether or not the underlying satellite DNA phylogenies were robust under bootstrap analysis.

Introduction

Satellite DNAs are tandemly repeated DNA sequences that differ from micro- and mini-satellite DNAs on account of their large size (two to several thousand bp), high copy number (10^3 – 10^7 at each locus), the small number of loci they represent, and location primarily in the centromeric heterochromatin (Tautz 1993). Current research focuses on their structural and functional relevance (Koch 2000; Lee et al. 2000), the underlying evolutionary processes (Lee et al. 1997; Nabeyama et al. 2000) and their use as taxonomic (King and Cummings 1997; Fedorov et al. 1999) and phylogenetic markers (Árnason et al. 1992; Garrido-Ramos et al. 1999). Available data are mostly sequences for single, i.e., monomeric units. Attempts have been made to elucidate the higher-

order organization of tandemly repeated satellites through the study of multimers and long sequences generated with restriction enzymes and the polymerase chain reaction. However, progress may be hampered due to flawed classifications and the absence of quantitative tests. I here show that phylogenetic network analysis improves the existing classification of satellite DNAs found in *Triturus* salamanders, and that the new classification increases the testability of hypotheses on satellite DNA evolution.

Up to 10% of the genome of *Triturus* aquatic salamanders is composed of a 32–33 bp ‘TkS1’ satellite DNA, corresponding to 60 million copies per haploid genome. A convenience or ad hoc classification (sensu Wiley 1981, p 197) yielded three ‘types’ of TkS1 (Varley et al. 1990). In contrast, the classification of the satellite DNAs derived under the parsimony criterion yields four types and is non-ambiguous. It gives rise to a hypothesis on the evolutionary history of TkS1 with high information content. Specifically, I uncovered statistically significant support for particular repetitive motifs, conforming to phylogenetic considerations at the species level. Satellite DNAs from *Palorus* beetles analyzed for comparison also provide statistically relevant support for the existence of alternating repetitive motifs.

Materials and methods

Triturus salamanders

The material analyzed consists of 244 satellite DNAs that were sequenced from cloned fragments of genomic DNA in five species of the monophyletic *Triturus cristatus*-*Triturus marmoratus* species group. Digestion with HaeIII yielded 207 32–33 bp monomers and nine (ca 64 bp) dimers. Treatment with AluI yielded 20 monomers, 15 dimers and two (ca 96 bp) trimers. Aligned sequences are numbered 1–161 in published order including synonyms (Varley et al. 1990: Tables 3–5, 7–13). In network analysis alignment gaps were coded as nucleotides that were absent at that particular position. Accordingly, gaps at position 14 are coded D and a gap at position 11 is coded G. Additionally, two gaps at position 29 are coded M. The artificial homoplasmy among haplotypes 14, 38, 52 and 104 does not affect my interpretations. The AluI 24–25 bp ‘heads’ and 8 bp ‘tails’ (in multimers flanking full TkS1

Edited by: W. Hennig

J.W. Arntzen (✉)
National Museum of Natural History, P.O. Box 9517,
2300 RA Leiden, The Netherlands
e-mail: arntzen@nmm.nl

J.W. Arntzen
Centro de Investigação em Biodiversidade e Recursos Genéticos,
Campus Agrário de Vairão, Rua do Monte, Crasto,
4485-661 Vairão, Portugal

sequences) were analyzed a posteriori, when Tks1 classification criteria had been formulated.

Palorus beetles

The material analyzed consists of seven 'PFIC' dimers from *Palorus ficicola*, six 'PGEN' trimers from *Palorus genalis* and seven 'PSUB' dimers from *Palorus subdepressus* (Plohl et al. 1998; Fig. 1; Mestrovic et al. 2000: Figs. 2, 3). Monomer size is 72 bp for PFIC and PSUB and ca 144 bp for PGEN. Missing data (5.0% in PFIC) were filled in with consensus sequence information. Gaps were dealt with in network analysis as described for the salamanders.

Calculations

Sequences were homologized and aligned as published and interpreted as haplotypes. Phylogenetic relationships of haplotypes were initially analyzed with PAUP 3.1 software (Swofford 1993) under default parameter settings. Internal branches of zero length were collapsed to yield polytomies. Phylogenetic information was summarized by the computation of a strict consensus tree. Bootstrap replication scores were determined from 2000 bootstrap replicates, if necessary using the 'fast addition' option in PAUP 4. The constraint of a non-reticulated branching structure was relaxed in the construction of median-joining networks with NETWORK 2.0 software under default parameter settings (Röhl 1997). Measures of gene diversity and nucleotide diversity were calculated with AM-

OVA software (Schneider et al. 1997). Mantel tests were used to test the hypothesis that positional dissimilarity (e.g., dimer heads versus tails) is associated with sequence dissimilarity, as counted over the network, with NTSYS 1.50 (Rohlf 1988). *G*-tests and other statistical analyses followed Sokal and Rohlf (1981).

Results

Phylogenetic analysis of 113 non-synonymous Tks1 haplotypes yielded 32.767 (the maximum storage capacity) unrooted trees of 138 steps. The strict consensus tree is shown schematically in Fig. 1 (bottom insert). Five clusters of haplotypes (1, 2a, 2b, 3 and 4) are separated from one another by two or more substitutions, with low bootstrap support (<0.50 in all cases). The classification is unambiguous, except for haplotype 113, which is here arbitrarily allocated to type 4 and not type 3.

The median network accommodates 244 haplotypes (of which 113 are non-synonymous) and has seven clusters, each with several unconnected branches sprouting from the corresponding nodes (Fig. 1). Six hypothetical haplotypes are inferred at nodes that have no unconnected branches. The network resolves the same types as the tree. One inferred haplotype (involving two character

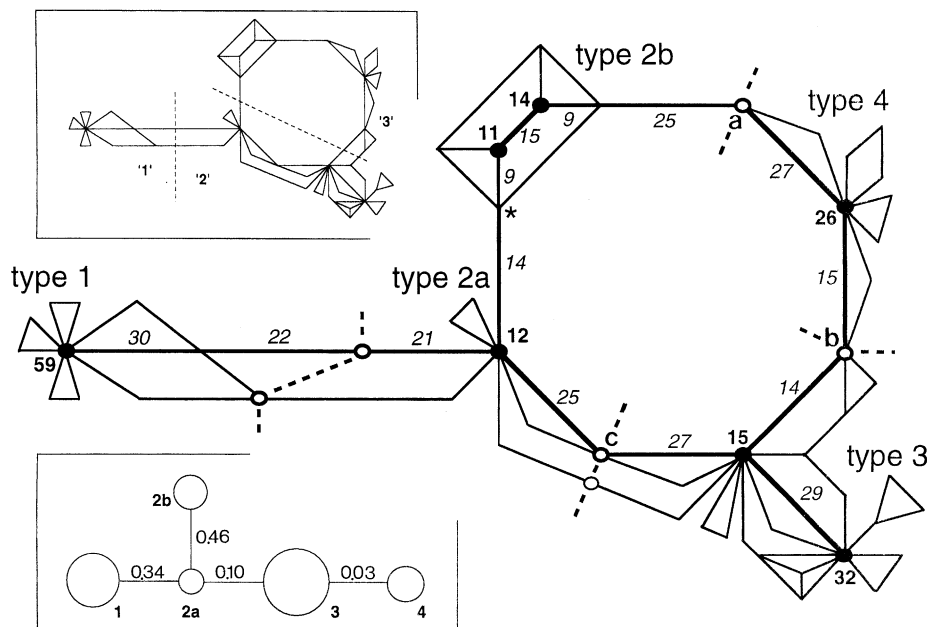


Fig. 1 Median network for 244 Tks1 satellite DNA sequences (haplotypes) from *Triturus aquatic* salamanders. *Solid dots* represent commonly observed haplotypes ($n \geq 5$), with n plus the number of unconnected branches (not shown for reasons of clarity) sprouting from the corresponding node in *bold*. *Open dots* represent hypothetical haplotypes at nodes with no unconnected branches. Dots are connected by a *thick line* with nucleotide positions subject to character state change in *italics*. Haplotypes are classified into types 1–4. At least two character state changes and one hypothetical haplotype separate types from one another, as highlighted by *interrupted lines*. Diagnostic character states are listed in the text. Types 2a and 2b are connected through an observed haplotype at the node marked with an *asterisk*. The letters *a*, *b* and *c* in-

dicate unobserved haplotypes at which the ring in the network can be broken to obtain the following type unreticulated arrangements: *a* 1-2-3-4; *b* 1-2 forking to 3 and 4; *c* 1-2-4-3. *Top insert* Classification by Varley et al. (1990) superimposed on the network. Note that 'type 1' is largely congruent between classifications. *Bottom insert* Schematic representation of the strict consensus tree of 32.767 equally parsimonious bifurcating trees (113 non-synonymous Tks1 haplotypes in analysis). *Circles* proportional in size to the frequency of occurrence represent the five clusters of haplotypes. Branches do have low bootstrap replication (<0.50), but do not collapse in consensus trees for subsets of haplotypes (1-2a, 2a-b, 2a-3, 2b-4)

Table 1 Observed distribution of four types of TkS1 satellite DNA over five species of salamanders (genus *Triturus*) with measures of genetic diversity

Species and individuals	Restriction enzyme employed	Type				Type 1+4 combined frequency
		1	2	3	4	
<i>T. carnifex</i>	HaeIII	3	9	11	1	0.17
<i>T. cristatus</i>	HaeIII	11	15	13	2	0.32
<i>T. dobrogicus</i>	HaeIII	2	3	18	2	0.16
<i>T. karelinii</i> no. 1	HaeIII	21	9	11	12	0.62
<i>T. karelinii</i> no. 1	AluI	1	0	5	13	^a
<i>T. karelinii</i> no. 2	HaeIII	23	5	7	3	0.68
<i>T. karelinii</i> no. 3	HaeIII	15	1	6	5	0.74
<i>T. marmoratus</i>	HaeIII	0	12	5	0	0.00
Gene diversity		0.634	0.912	0.904	0.804	
Standard deviation		0.066	0.028	0.028	0.068	
Nucleotide diversity		0.033	0.078	0.060	0.043	
Standard deviation		0.024	0.048	0.038	0.030	

^a Not applicable due to constraints posed by the restriction enzyme used

Table 2 Probability of the observed bi/tri-partitioning of haplotypes over the network following (i) the Pascal triangle, (ii) Mantel permutation tests for association between sequence similarity

Taxon	Satellite DNA	Haplotypes	Pascal	Mantel	Bootstrap replication score
<i>P. ficicola</i>	PFIC	A and B	$P < 0.05$	$P < 0.001$	0.29
<i>P. genalis</i>	PGEN	A, B and C	$P < 0.01$	$P < 0.001$	1.00, 1.00
<i>P. subdepressus</i>	PSUB	A and B	$P > 0.05$	$P < 0.01$	0.67

and positional similarity and (iii) bootstrap replication scores over maximum parsimony, non-reticulated phylogenetic trees derived for satellite DNAs from three species of *Palorus* flour beetles

state changes) separate the types from one another. However, haplotype 79 is intermediate between groups 2a and 2b and hence remains unclassified at that level. Diagnostic character states at positions 9, 14, 15, 21, 22, 25, 27, 29 and 30 are as follows: type 1, A-TAGGTTG; type 2a, A-TCTGTTT; type 2b, CCTCTGTTT; type 3, A-TCTACKT; and type 4, ACCCTACTT. The limited variation resolved from AluI digests at positions 1, 2, 32 and 33, which make up the HaeIII recognition site, does not contribute to a sharper or deeper classification of TkS1. It is important to note that AluI recognition sequences (AGCT) cover positions 25 and 27, which are diagnostic for types 1 and 2 versus types 3 and 4. In consequence, AluI-derived monomers have type 1 or type 2 (and not type 3 or type 4) heads and tails, while AluI-derived multimers have type 1 or 2 heads and tails and type 3 or 4 'bodies'.

The unrooted tree and network are highly congruent, with the main difference that type 2 and type 4 in the tree are not directly connected while in the network they are. The character states inferred to be homoplasous are at nucleotide positions 14 and 15. Haplotype 72 is classified as type 3 in the network and as type 2a in the tree. I here follow the latter option.

Gene diversity is lower in type 1 and type 4 than in type 2 and type 3 ($P < 0.05$). Nucleotide diversity shows the same trend but differences between types are not significant ($P > 0.05$; Table 1). No significant differences are observed in the distribution of types over three *T. karelinii* individuals (G -test of independence, $G=12.8$, $df=8$, $P > 0.05$). However, the distribution of types over individuals from other species is not random ($G=85.3$,

$df=16$, $P < 0.01$), with the combined frequency of type 1 and type 4 increasing in the order *T. marmoratus* (0%), *T. dobrogicus* (16%), *T. carnifex* (17%), *T. cristatus* (32%) and *T. karelinii* (67%). The association of heads and tails in 20 AluI monomers is not random (type 1-type 1, $n=12$; type 1-type 2, $n=3$; type 2-type 1, $n=0$; type 2-type 2, $n=5$; $P < 0.01$, Fisher's exact test), suggesting that haplotypes are mostly homosequentially arranged. Similarly, the association of heads, bodies and tails in 15 AluI dimers is not random (type 4 flanked by type 1, $n=12$; type 3 flanked by type 2, $n=3$; other combinations, $n=0$; $P < 0.01$, Fisher's exact test), suggesting that heterosequential tandem repeats are mostly 1-4 and 2-3 combinations. Predictions on type composition of tandem repeats were confirmed from nine HaeIII dimers. Four dimers were homosequential (1-1, 2-2 and 3-3), two were 1-4 and three were 2-3 heterosequential combinations ($P < 0.05$ against type frequencies as observed in the specimen of *T. karelinii* from which dimers were taken, Table 1; $P < 0.01$ against frequencies across three *T. karelinii* individuals; binomial tests).

Palorus beetles

Medium networks of haplotypes in *P. ficicola*, *P. genalis* and *P. subdepressus* show reticulation to a larger or smaller extent (Figs. 2, 3, 4). Haplotypes that share head, body or tail position within dimers and trimers are significantly clustered over the networks in *P. ficicola* and *P. genalis* and not in *P. subdepressus* (Table 2). Mantel tests support the hypothesis that haplotypes sharing posi-

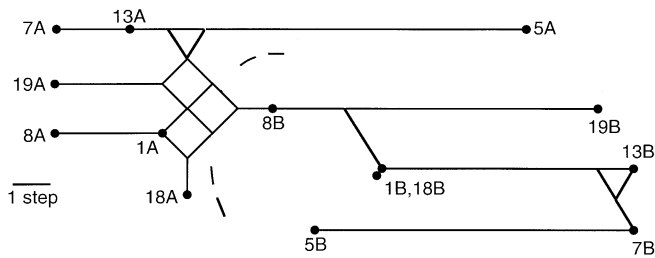
Palorus ficicola

Fig. 2 Median network for 14 PFIC satellite DNA sequences from the flour beetle *Palorus ficicola*. Observed haplotypes are shown by dots and numbered as in the source publication (Meštrović et al. 2000). The interrupted line separates haplotype groups. A and B refer to the position of the haplotype within a dimer

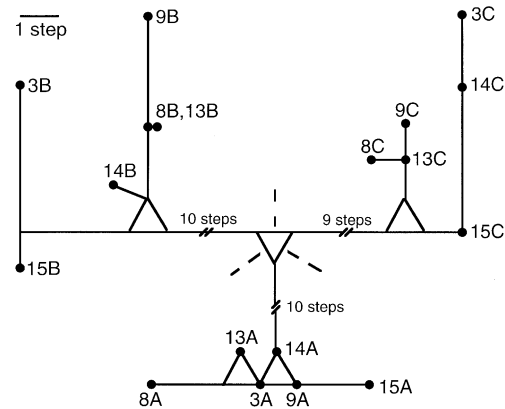
Palorus genalis

Fig. 3 Median network for 18 PGEN satellite DNA sequences from the flour beetle *Palorus genalis*. A, B and C refer to the position of the haplotype within a trimer. Otherwise as in Fig. 2 legend

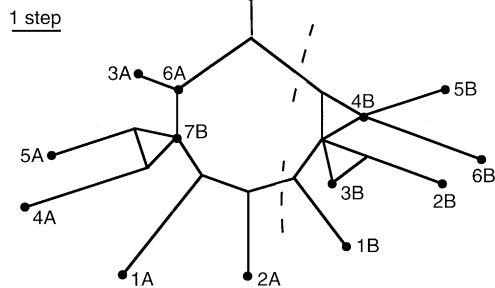
Palorus subdepressus

Fig. 4 Median network for 14 PSUB satellite DNA sequences from the flour beetle *Palorus subdepressus* (Pohl et al. 1998). Otherwise as in Fig. 2 legend

tion are more similar in sequence composition than are haplotypes that take different positions. Bootstrap replication scores for the diagnostic branches in maximum parsimony (non-reticulated) trees range from 0.29 to 1.00.

Discussion

I analyzed published Tks1 satellite DNA sequences under the principle of parsimony. This resulted in an essentially unambiguous classification of the satellite DNAs in four types. Conversely, the published classification, here represented in Fig. 1 (top insert), is highly ambiguous and does not necessarily bear a relationship to the pattern and process of evolutionary change, as is admitted by the authors (Varley et al. 1990).

The phylogenetic network connects the four types in a 'p'-shaped arrangement whereas under the constraint of non-reticulation the arrangement is 'l'- or 'y'-shaped. In accordance with the criteria formulated for the recognition of types the ring in the network can be broken at one of three positions representing unobserved haplotypes (marked a, b and c in Fig. 1). This gives rise to the following unreticulated arrangements (a) 1-2-3-4; (b) 1-2 forking to 3 and 4; (c) 1-2-4-3. Two lines of evidence support the first of these alternatives by suggesting peripheral positions for types 1 and 4 and central positions for types 2 and 3. Firstly, the inferred absence of types 1 and 4 in *T. marmoratus* versus their presence in its sister taxon (the *T. cristatus* superspecies, comprising *T. carnifex*, *T. cristatus*, *T. dobrogicus* and *T. karelinii*; Arntzen and Wallis 1999) suggests that these are younger than types 2 and 3. Secondly, genetic diversity is lower in the former than in the latter, as would be predicted for more recently evolved types. This inference also renders the 1-4 combination a shared derived character state for the *T. cristatus* superspecies.

Dimeric and pentameric higher order structures are well established for the α -satellite DNAs in the primate genome (Haaf and Willard 1997; Loftus et al. 1999). The dimeric and trimeric alternating motifs reported earlier for *Palorus* beetles (Pohl et al. 1998; Meštrović et al. 2000) are here shown to have statistical support. I furthermore uncovered support for the presence of homosequential tandem repeats (type 1-1, type 2-2), as well as for type 1-4 and type 2-3 and not other heterosequential tandem repeats in *Triturus* salamanders. Types 1 and 4 have proliferated in *T. karelinii* in particular and they appear to have done so as a tandemly repeated unit. Various molecular genetic processes may account for the ordering (slip-page, unequal crossover, rolling circle replication processes complicated by recombination; Walsh 1987; Rossi et al. 1990; Jeffreys et al. 1998). Modeling indicates that these mechanisms will be difficult to distinguish in practice (Stephan and Cho 1994). This will be particularly true if the perceived higher order structure is based upon a flawed classification, such as the trimeric Tks1 '1-3-1 supertype' of Varley et al. (1990).

The phylogenetic analysis of repeats will often be a challenging task because of the large number of samples and small differences between them. The resulting multitude of plausible solutions is best expressed by a network, which displays alternative potential evolutionary paths in the form of cycles (Bandelt et al. 1999). A high level of reticulation will be reflected in a low level of nu-

merical support for the corresponding branching structure (such as expressed by a bootstrap replication score, Felsenstein 1985; Hillis and Bull 1993). I point out that it is the final hypothesis that requires statistically significant support and not any of the constituent constructs. Dense reticulations and low bootstrap values as found for *Triturus* Tks1 and *Palorus* PFIC and PSUB phylogenetic trees and networks reflect the complexity of the data. They do not compromise the support for higher order repeat organization.

Multicopy 'satellite' DNA is found in a wide variety of organisms. Most of it is transcriptionally inactive. Whether or not satellite DNAs are functional is under discussion, but surely their high copy number provokes doubts as to their selective neutrality and raises questions about their evolutionary significance. Several roles have been postulated, mostly associated with centromeric function (e.g., Ugarković et al. 1996; Capriglione et al. 1998; Koch 2000). Other authors merely find it difficult to concede that the large fractions of genomes that satellite DNAs constitute have little or no phenotypic consequences (see Elder and Turner 1995 for a review; for a discussion on the functional roles of microsatellite and minisatellite DNAs see Kashi and Soller 1999). I anticipate that testable and informative hypotheses on the origin, structural organization, mechanisms of proliferation, behavior and – ultimately – function of multicopy satellite DNAs will have their roots in well-designed classifications.

References

- Árnason Ú, Grétarsdóttir S, Widegren B (1992) Mysticete (baleen whale) relationships based upon the sequence of the common cetacean DNA satellite. *Mol Biol Evol* 9:1018–1028
- Arntzen JW, Wallis GP (1999) Geographic variation and taxonomy of crested newts (*Triturus cristatus* superspecies): morphological and mitochondrial DNA data. *Contrib Zool* 68:181–202
- Bandelt H-J, Forster P, Röhl A (1999) Median-joining networks for inferring intraspecific phylogenies. *Mol Biol Evol* 16:37–48
- Capriglione T, Santo MG de, Odierna G, Olmo E (1998) An alpha-oid-like satellite DNA sequence is present in the genome of a lacertid lizard. *J Mol Evol* 46:240–244
- Elder JF, Turner BJ (1995) Concerted evolution of repetitive DNA sequences in eukaryotes. *Quart Rev Biol* 70:297–320
- Fedorov AN, Fedorova LV, Grechko VV, Ryabinin DM, Sheremet'eva VA, Bannikova AA, Lomov AA, Ryskov AP, Darevsky IS (1999) Variable and invariable DNA repeat characters revealed by taxonprint approach are useful for molecular systematics. *J Mol Evol* 48:69–76
- Felsenstein J (1985) Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* 39:783–791
- Garrido-Ramos MA, Herrán R de la, Jamilena M, Lozano R, Ruiz Rejón C, Ruiz Rejón M (1999) Evolution of centromeric satellite DNA and its use in phylogenetic studies of the Sparidae family (Pisces, Perciformes). *Mol Phylogenet Evol* 12:200–204
- Haaf T, Willard HF (1997) Chromosome-specific α -satellite DNA from the centromere of chimpanzee chromosome 4. *Chromosoma* 106:226–232
- Hillis DM, Bull JJ (1993) An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. *Syst Biol* 42:182–192
- Jeffreys AJ, Neil DL, Neumann R (1998) Repeat instability at human minisatellites arising from meiotic recombination. *EMBO J* 17:4147–4157
- Kashi Y, Soller M (1999) Functional roles of microsatellites and minisatellites. In: Goldstein DB, Schlötterer C (eds) *Microsatellites. Evolution and Applications*. Oxford University Press, Oxford
- King LM, Cummings MP (1997) Satellite DNA repeat sequence variation is low in three species of burying beetles in the genus *Nicrophorus* (Coleoptera: Silphidae). *Mol Biol Evol* 14:1088–1095
- Koch J (2000) Neocentromeres and alpha satellite: a proposed structural code for functional human centromere DNA. *Hum Mol Genet* 9:149–154
- Lee C, Court DR, Cho C, Haslett JL, Lin C-C (1997) Higher-order organization of subrepeats and the evolution of cervid satellite I DNA. *J Mol Evol* 44:327–335
- Lee C, Critcher R, Zhang J-G, Mills W, Farr CJ (2000) Distribution of gamma satellite DNA on the human X and Y chromosomes suggest that it is not required for mitotic centromere function. *Chromosoma* 109:381–389
- Loftus B, Kim U-J, Sneddon VP, Kalush F, Brandon R, Fuhrmann J, Mason T, Crosby ML, Barnstead M, Cronin L, Deslattes Mays A, Cao Y, Xu RX, Kang H-L, Mitchell S, Eichler EE, Harris PC, Venter JC, Adams MD (1999) Genome duplications and other features in 12 Mb of DNA sequence from human chromosome 16p and 16q. *Genomics* 60:295–308
- Meštrović N, Mravinac B, Juan C, Ugarković D, Plohl M (2000) Comparative study of satellite sequences and phylogeny of five species from the genus *Palorus* (Insecta, Coleoptera). *Genome* 43:776–785
- Nabeyama M, Kubota S, Kohno S (2000) Concerted evolution of a highly repetitive DNA family in Eptatretidae (Cyclostomata, Agnatha) implies specifically differential homogenization and amplification events in their germ cells. *J Mol Evol* 50:154–169
- Plohl M, Meštrović N, Bruvo B, Ugarković D (1998) Similarity of structural features and evolution of satellite DNAs from *Palorus subdepressus* (Coleoptera) and related species. *J Mol Evol* 46:234–239
- Röhl A (1997) Network 2.0b. A program package for phylogenetic networks. Mathematisches Seminar, Universität Hamburg. Available from <http://www.fluxus-engineering.com>
- Rohlf FJ (1988) NTSYS-1.50. Numerical taxonomy and multivariate analysis system. Computer program. Exeter, New York
- Rossi MS, Reig OA, Zorzopulos J (1990) Evidence for rolling-circle replication in a major satellite DNA from the South American rodents of the genus *Ctenomys*. *Mol Biol Evol* 7:340–350
- Schneider S, Kueffer J-M, Roessler D, Excoffier L (1997) Arlequin 1.1. A software for population genetic data analysis. Genetics and Biometry Laboratory, University of Geneva, Geneva
- Sokal RR, Rohlf FJ (1981) *Biometry*, 2nd edn. Freeman, San Francisco
- Stephan W, Cho S (1994) Possible role of natural selection in the formation of tandem-repetitive noncoding DNA. *Genetics* 136:333–341
- Swofford DL (1993) PAUP: phylogenetic analysis using parsimony, version 3.1. Computer program. Distributed by Illinois Natural History Survey, Champaign, Illinois
- Tautz D (1993) Notes on the definition and nomenclature of tandemly repetitive DNA sequences. In: Pena SDJ, Chakraborty R, Eppel JT, Jeffreys AJ (eds) *DNA fingerprinting: state of the science*. Birkhäuser, Basel
- Ugarković D, Podnar M, Plohl M (1996) Satellite DNA of the red flour beetle *Tribolium castaneum* – comparative study of satellites from the genus *Tribolium*. *Mol Biol Evol* 13:1059–1066
- Varley JM, Macgregor HC, Barnett L (1990) Characterisation of a short, highly repeated and centromerically localised DNA sequence in crested and marbled newts of the genus *Triturus*. *Chromosoma* 100:15–31
- Walsh JB (1987) Persistence of tandem arrays: implications for satellite and simple-sequence DNAs. *Genetics* 115:553–567
- Wiley EO (1981) *Phylogenetics. The theory and practice of phylogenetic systematics*. Wiley, New York