Slipped-Strand Mispairing at Noncontiguous Repeats in *Poecilia reticulata*: A Model for Minisatellite Birth

John S. Taylor and Felix Breden

Department of Biological Sciences, Simon Fraser University, British Columbia, V5A 1S6, Canada

Manuscript received August 10, 1999
Accepted for publication March 16, 2000

ABSTRACT

The standard slipped-strand mispairing (SSM) model for the formation of variable number tandem repeats (VNTRs) proposes that a few tandem repeats, produced by chance mutations, provide the "raw material" for VNTR expansion. However, this model is unlikely to explain the formation of VNTRs with long motifs (*e.g.*, minisatellites), because the likelihood of a tandem repeat forming by chance decreases rapidly as the length of the repeat motif increases. Phylogenetic reconstruction of the birth of a mitochondrial (mt) DNA minisatellite in guppies suggests that VNTRs with long motifs can form as a consequence of SSM at noncontiguous repeats. VNTRs formed in this manner have motifs longer than the noncontiguous repeat originally formed by chance and are flanked by one unit of the original, noncontiguous repeat. SSM at noncontiguous repeats can therefore explain the birth of VNTRs with long motifs and the "imperfect" or "short direct" repeats frequently observed adjacent to both mtDNA and nuclear VNTRs.

VARIABLE number tandem repeats (VNTRs), especially microsatellites, have become the genetic marker of choice for forensics (Hagelberg *et al.* 1991; Olaisen *et al.* 1997), genomic mapping (Dib *et al.* 1996; Dietrich *et al.* 1996), and quantifying intraspecific genetic variation (Bowcock *et al.* 1994; Schlötterer *et al.* 1997; Lunt *et al.* 1998). Furthermore, many human genetic diseases are caused by microsatellite VNTR expansion (Mandel 1997) and an increase in VNTR variability is an indicator of mutations associated with several forms of cancer (Wada *et al.* 1994). Despite the ubiquity and importance of VNTRs, little is known about the molecular mechanisms leading to their formation.

This lack of empirical data on repeat formation is surprising, because VNTR formation must be a frequent occurrence. This conclusion is based on the observation that, while all eukaryotic taxa surveyed possess VNTRs, studies where several species have been tested with the same VNTR primers reveal that a given repeat rarely occurs in more than a few closely related taxa. In some of these studies a repeat that is perfect and highly variable in one taxon is interrupted or very short and monomorphic in close relatives (Zardoya *et al.* 1996; Brohede and Ellegren 1999). In many other studies, variable repeats in one taxon are unrecognizable in close relatives (Blanquer-Maumont and Crouau-Roy 1995; Angers and Bernatchez 1997; Taylor *et al.* 1999).

Slipped-strand mispairing (SSM; Levinson and Gutman 1987) is the most often proposed model for repeat formation, expansion, and contraction. A critical component of the SSM model for repeat formation is the occurrence of chance mutations that produce a few tandem repeats facilitating the first strand slippage event. Levinson and Gutman (1987) refer to these tandem repeats produced by chance mutations as the "raw material" for repeat expansion by SSM. A study of repeat evolution in primates demonstrated that chance mutations played a role in the formation of two VNTRs, one with a 2-bp motif and one with a 4-bp motif (Messier *et al.* 1996). By mapping sequence data onto a phylogeny, Messier *et al.* (1996) discovered that in the owl monkey an A-to-G mutation produced $(GT)_5$, which expanded, presumably via SSM, to $(GT)_6$. Messier *et al.* (1996) also discovered that a G-to-A mutation produced $(ATGT)_2$ in Hominoidea, which expanded to $(ATGT)_4$ in gorillas, bonobos, and chimpanzees and to $(ATGT)_5$ in humans. Thus, the model proposed by Levinson and Gutman (1987) is supported by the microsatellite evolution data reported by Messier *et al.* (1996). However, it is unknown whether these two instances of VNTR birth are typical, or whether chance mutations are important in the formation of repeats with motifs longer than 4 bp.

While surveying guppy (*Poecilia reticulata*) populations for mtDNA control region sequence variation we uncovered a minisatellite with an 11-bp motif in individuals from a tributary of the Rio Grande in Trinidad. We investigated the evolution of this minisatellite by mapping mutations onto a population-level phylogeny. This led to a general hypothesis that SSM at noncontiguous repeats can lead to the birth of VNTRs with long motifs and characteristic flanking repeats.

Corresponding author: John S. Taylor, Department of Biology, University of Konstanz, D-78457 Konstanz, Germany.
E-mail: john.taylor@uni-konstanz.de

TABLE 1

Guppy (*P. reticulata*) populations surveyed for mitochondrial control region sequence variation

Symbol	Location	Individuals sequenced		GenBank accession nos.
		911 bp	150 bp[a]	
	Trinidad			
ACR	Arima River at Churchill-Roosevelt Hwy.	ACR9		AF170264
TrT	Arima River at Tripp Trace		TrT1	AF170266
ASA	Arima River near Asa Wright		ASA1	AF228623
APL	Aripo River at Eastern Main Rd.		APL182	AF080489
RAP	Aripo River at Rapsey's Farm		RAP3	AF170268
GUA	Guanapo River		GUA1	AF170267
OVR	Oropuche River at Valencia Main Rd.	OVR6		AF193899
OCR	Oropuche River at Cumaca Rd.		OCR3	AF170259
NOR	Oropuche River at Norbert's Ranch		NOR1	AF170260
QU4	Quare River at Water Works	QU48		AF193897
QU6	Quare River tributary (bridge before dam)		QU6F1	AF170261
QMD	Quare River: "Mike's downstream site"		QMD1	AF193898
RG	Tributary of Rio Grande	RG1	RG4, 5, 6	AF170269, AF170258,70–71
AQI	Aqui River (tributary of Madamas River)		AQI3	AF170262
PAU97	Paria River at Brasso Seco		PAU971	AF193902
JOR	Jordan River: Tributary of Paria River		JOR1	AF228624
MAU	Marianne River at Brasso Seco Road		MAU0049	AF193901
YAR	Yarra River at Maracas Royal Road	YAR1		AF170265
YAM, LIM	Yarra River at Malmoral Trace		YAM1, LIM1	AF228625, AF170263
	Venezuela			
MAR	Rio El Valle, Isla de Margarita		MAR3	AF228610
BTB	Guanare River at El Puente	BTB3	BTB4	AF170257, AF228614
SUS	Guanare River at Anzoátegui		SUS3	AF228615
ElT	El Tacque		ElT3, 6	AF228616–17
ME	Rio Medio		ME3, 4, 5	AF228618–20
PV	Poza de Azufre (Sulphur Spring)		PV4	AF228621
PV6	River 6 Km from Pozo Azufre		PV61	AF228622
VP	Orinoco River, Delta Amacuro		VP1, 2, 3	AF228611–13
LaC	Las Claritas		LaC2	AF170254
	Guyana			
BAR	Essequibo River (Bartica town trenches)	BAR3		AF170255
CHA	Pomeroon River (Charity town trenches)		CHA1, 2	AF170256, AF228604
NWA	New Amsterdam town trenches		NWA3	AF228609
SPL	Springlands town trenches		SPL3	AF228608
	Surinam			
LYD	Lelydorp, home of John DeBruin		LYD3, 4, 5	AF228605–07
Total		7	39	

[a] The 150-bp fragment includes the minisatellite and is included in the 911-bp fragment.

MATERIALS AND METHODS

Sampling and molecular techniques: Our survey included 46 guppies from 33 sites in Trinidad, Venezuela, Guyana, and Surinam (Table 1). In the field, guppies were preserved in 95% ethanol. In the lab, DNA was isolated from the tail musculature using methods described by Fajen and Breden (1992). To produce sequencing templates, we amplified ~1000 bp of mitochondrial control region DNA using the primers L15926 (Kocher *et al.* 1989) and MRT2 (Ptacek and Breden 1998). PCR conditions included an initial denaturation at 94° for 1 min, followed by 35 cycles each consisting of denaturation at 94° for 1 min, annealing at 52° for 1 min 20 sec, and extension for 2 min at 72°. PCR products were purified on a 1% agarose gel. A small block of agarose containing the PCR product was cut out of the gel and frozen overnight. This gel block was then spun in a microcentrifuge at high speed for 7 min. Two microliters of the resulting liquid was used as a sequencing template. For 39 samples from 26 locations (Table 1), we sequenced both strands of the left domain or R1 portion

TABLE 2

Partial mitochondrial control region sequences for 46 guppies (*P. reticulata*)

Sample	Repeat-containing portion of the mitochondrial control region
OVR6	TTGACCAAAATCTGCCCCAAATAT.....ATGTACT-TTTATAGT
OCR3	TTGACCAAAATCTGCCCCAAATAT.....ATGTACT-TTTATAGT
NOR1	TTGACCAAAATCTGCCCCAAATAT.....ATGTACT-TTTATAGT
AQI3	TTGACCAAAATCTGCCCCAAATAT.....ATGTACT-TTTATAGT
QU48	TTGACCAAAATCTGCCCCAAATAT.....ATGTACT-TTTATAGT
QU6F1	TTGACCAAAATCTGCCCCAAATAT.....ATGTACT-TTTATAGT
QMD1	TTGACCAAAATCTGCCCCAAATAT.....ATGTACT-TTTATAGT
RG1	TTGACCAAAATCTGCcccaaatct.....ATGTACT-TTTATAGT
RG4	TTGACCAAAATCTGCcccaaatct.....ATGTACT-TTTATAGT
RG5	TTGCCCAAAATCTGCCCCAAATCC.....ATGTACT-TTTATAGT
RG6	TTGACCAAAATCTGCcccaaatct.....ATGTACT-TTTATAGT
BAR3	TTGACCAAAATCTGCCCCAAATAC.....ATGTACTATTTATAGT
CHA1	TTGACCAAAATCTGCCCCAAATAC.....ATGTACTATAAATAGT
CHA2	TTGACCAAAATCTGCCCCAAATAC.....ATGTACTATAAATAGT
LaC2	TTGGCCAAAATCTGCCCCAAATAC.....ATGTAC-ATTTATAGT
LYD3	TTGGCCAAAATCTGCCCCAAAAAC..CTGCCCCAAATGTACTATGTATAGT
LYD4	TTGGCCAAAATCTGCCCCAAAAAC..CTGCCCCAAATGTACTATGTATAGT
LYD5	TTGGCCAAAATCTGCCCCAAAAAC..CTGCCCCAAATGTACTATGTATAGT
SPL3	TTGGCCAAAATCTGCCCCAAAAAT..CTGCCCCAAATGTACTATAAATAGT
NWA3	TTGGCCAAAATCTGCCCCAAAAAT..CTGCCCCAAATGTACTATAAATAGT
MAR3	TTGACCGAAATCTGCCCCAAATAT.....ATGTACTATTTATAGT
VP1	TTGACCGAAATCTGCCCCGAATAT.....ATGTACTATTTATAGT
VP2	TTGACCGAAATCTGCCC-TAATAT.....ATGTAC-ATTTATAGT
VP3	TTGACCGAAATCTGCCCCGAATAT.....ATGTACTATTTATAGT
BTB3	TTGACCGAAATCTGCCC-TAATAT.....ATGTACCATTTATAGT
BTB4	TTGACCGAAATCTGCCC-TAATAT.....ATGTACCATTTATAGT
SUS3	TTGACCGAAATCTGCCC-TAATATAT.....ATGTACTATTTATAGT
ElT3	TTGACCGAAATCTGCCC-TAATATAT.....ATGTACTATTTATAGT
ElT6	TTGACCGAAATCTGCCC-TAATATAT.....ATGTACTATTTATAGT
ME3	TTGACCGAAATCTGCCCCGAATAT.....ATGTACTATTTATAGT
ME4	TTGACCGAAATCTGCCCCGAATAT.....ATGTACTATTTATAGT
ME5	TTGACCGAAATCTGCCCCGAATAT.....ATGTACTATTTATAGT
PV4	TTGACCGAAATCTGCCCCGAATAT.....ATGTACTATTTATAGT
PV61	TTGGCCGAAATCTGCCCCGAATAT.....ATGTACTATTTATAGT
APL182	TTGACCGAAATCTGCCCCGAATAT.....ATGTACTATTTATAGT
RAP3	TTGACCGAAATCTGCCCCGAATAT.....ATGTACTATTTATAGT
ACR9	TTGACCGAAATCTGCCCCGAATAT.....ATGTACTATTTATAGT
ASA1	TTGACCGAAATCTGCCCCGAATAT.....ATGTACTATTTATAGT
TrT1	TTGACCGAAATCTGCCCCGAATAT.....ATGTACTATTTATAGT
GUA1	TTGACCGAAATCTGCCCCGAATAT.....ATGTACTATTTATAGT
JOR1	TTGACCGAAATCTGCCCCGAATAT.....ATGTACTATTTATAGT
PAU971	TTGACCGAAATCTGCCCCGAATAT.....ATGTACTATTTATAGT
MAU0049	TTGACCGAAATCTGCCCCGAATAT.....ATGTACTATTTATAGT
YAR1	TTGACCGAAATCTGCCCCGAATAT.....ATGTACTATTTATAGT
LIM1	TTGACCGAAATCTGCCCCGAATAT.....ATGTACTATTTATAGT
YAM1	TTGACCGAAATCTGCCCCGAATAT.....ATGTACTATTTATAGT

The first three nucleotides are part of $tRNA^{PRO}$. Population symbols shown in Table 1. The repeat motif is underlined for this alignment. The RG1, RG4, and RG5 data were modified to include only one repeat motif. The 3′ partial repeat (see text) is shown in lowercase.

(Fumagalli *et al.* 1996) of the control region using primers L15995 (Meyer *et al.* 1994) and MR1 (5′-TAT GGG TTT TGT CTA CCT TC-3′). For seven individuals from across the guppy geographic distribution (Table 1), we sequenced 911 bp of the mitochondrial control region using the following primers, which were spaced ~200 bp apart: L15926, L15995, MR1, 12 RS (5′-CAT TTG GTT CCT ATT TCA GG-3′), 13 (5′-CAT TTC ACA GTG CAT ACA CA-3′), 14 (5′-AGT ATC CCC CTC GGC TTT TG-3′), 15 (5′-AAT TTT GTT TAC ATA CTT TA-3′), and MRT2. These primers provided overlapping sequences for 80–90% of the control region. Sequences from repeat-possessing Rio Grande (RG) guppies were unreadable beyond approximately three repeat units, suggesting that these samples were heteroplasmic (*i.e.*, possessed more than one mtDNA haplotype). To estimate the prevalence of the VNTR we end-labeled primer MR1 with $\gamma^{33}P$ and used MR1 and L15926 to

amplify the R1 portion of the control region in 18 RG samples. These 18 samples included the 4 samples sequenced for either 911 or 150 bp (Table 1). End-labeled PCR products were electrophoresed on a 6% acrylamide gel and visualized by exposure to X-ray film.

Analyses: Sequence alignments were performed using CLUSTAL V (Higgins *et al.* 1992). First, for all guppies ($N = 46$) the R1 portion (*ca.* 150 bp) of the control region was aligned. This alignment identified mutations occurring in the portion of the control region associated with repeat expansion. Second, complete control region sequences from seven *P. reticulata* samples and sequences from *P. caucana* (GenBank accession no. AF033057) and *P. parae* (accession no. AF033050) were aligned. Maximum parsimony analysis of this alignment produced a robust phylogeny of a subset of guppy populations upon which we mapped the changes in the R1 region. *P. caucana* occurs in the subgenus Poecilia and *P. parae* occurs in the subgenus Lebistes, along with the guppy (Breden *et al.* 1999). In this second alignment *P. reticulata* and *P. caucana* sequences were truncated so that they could be aligned with the shorter (813 bp) *P. parae* sequences. In both alignments Rio Grande sequences were modified to include only one repeat motif.

A maximum parsimony analysis was performed on complete control region sequences (*i.e.*, second alignment) using the heuristic search algorithm of PAUP version 3.1.1 (Swofford 1993). Deletions greater than one nucleotide long were treated as single characters in this analysis. PAUP settings included: addition sequence, stepwise (random seed number, 9); collapse zero-length branches; MULPARS option in effect. *P. caucana* was treated as the outgroup. Support for the tree topology was estimated using 500 bootstrap reiterations. *Xiphophorus nigrensis* sequence data (GenBank accession no. U06578) were added to the tree to polarize mutations.

RESULTS

Molecular results: Twenty individuals from 14 guppy populations possessed the 11-bp motif, CCAAAATC TGC, in the R1 portion of the control region (Table 2), but expansion of this motif was evident only in RG guppies. Deletions and duplications led to variation in the length of the R1 portion of the control region among the guppies surveyed (Table 2). Seventeen of the 18 RG samples surveyed using γ^{33}P-labeled MR1 and L15926 were heteroplasmic, possessing four to eight different-sized mtDNA haplotypes each.

Phylogenetic analyses: A bootstrap 50% majority-rule consensus tree is shown in Figure 1. The seven *P. reticulata* samples form a monophyletic group. Within the *P. reticulata* clade there are two groups with bootstrap values ≥99%. One includes the Yarra, Arima, and Guanare river samples. The sample from the Essequibo River (BAR3) appears to be the sister taxon to this clade. The second well-supported clade includes samples from the Oropuche Drainage in Trinidad (RG1, OVR6, and QU48).

A model for the birth of a minisatellite: By mapping sequence data onto the phylogeny, we uncovered two substitutions that appear to have been important for the formation of the minisatellite in RG guppies. First, the repeat motif is the consequence of a G-to-A mutation changing CCGAAATCTGC to CCAAAATCTGC. There are two equally parsimonious reconstructions of the changes at this site. First, this mutation may have occurred in the common ancestor of guppies from the Rio Grande, Quare, and Oropuche rivers and again in guppies from the Essequibo River (Figure 1). Alternatively, this mutation occurred in the common ancestor of all guppies surveyed with a reversal in the common ancestor of the Yarra River + Arima River + Guanare River clade. Our conclusion that the G is the ancestral

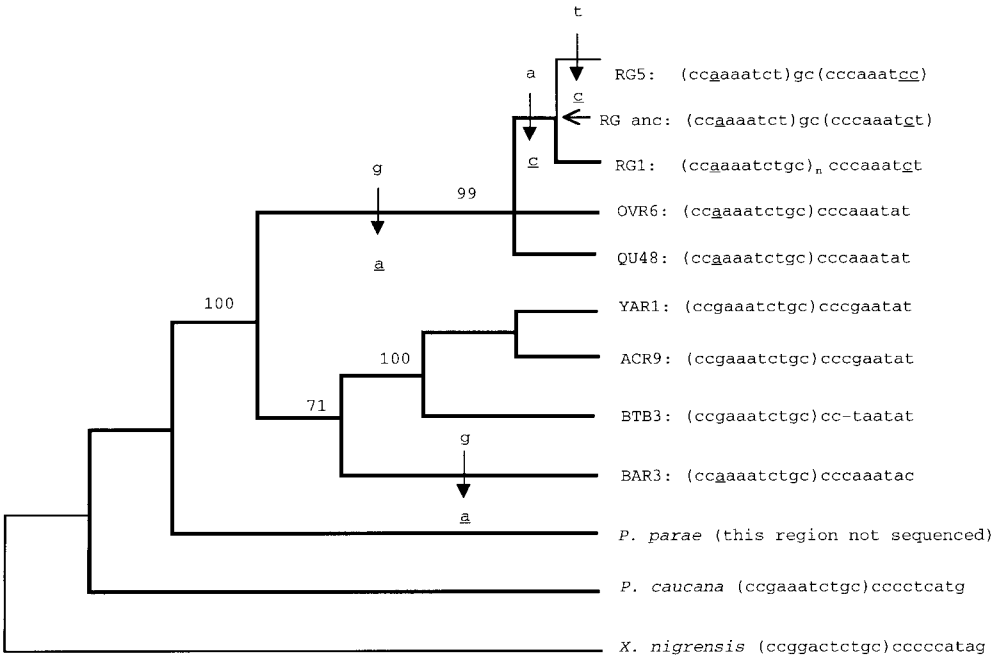

Figure 1.—Evolution of a noncontiguous imperfect repeat in guppies from the Rio Grande (RG). Mutations, including two that produce a noncontiguous repeat in the inferred ancestor of the RG guppies (RG anc), are represented by arrows on the phylogeny. Mutations mapped on the phylogeny involve the nucleotide sites that have been underlined in the control region sequences shown. The phylogeny is based upon maximum parsimony analysis of 817 bp of control region DNA. Numbers shown are bootstrap values (500 replications). Consistency index, 0.760; retention index, 0.785; homoplasy index, 0.240; rescaled consistency index, 0.597. RG5 was added after phylogenetic analyses based upon the A-to-C mutation it shares with other RG samples. Our model for repeat formation (Figure 2) does not depend upon the assumption of RG monophyly. *X. nigrensis* was added as an outgroup to Poecilia based upon ND2 sequence data (Breden *et al.* 1999) and was included only to polarize changes within the ingroup.

state for this nucleotide site is based upon the observations that CCGAAATCTGC occurs in guppies and *P. caucana* (Figure 1) and that CCGGACTCTGC occurs in *X. nigrensis*. The second substitution that appears to have been important for the formation of the RG minisatellite occurred in repeat flanking sequence. The guppies surveyed for mitochondrial sequence variation that have expanded repeats (RG1, 4, 6) possess a unique sequence, CCCAAATCT, adjacent to the repeat motif (Table 2; Figure 1). This repeat expansion-associated flanking sequence is apparently a consequence of an A-to-C mutation in the common ancestor of the RG guppies (changing CCCAAATAT to CCCAAATCT; Figure 1). We propose that the G-to-A and A-to-C mutations described above created an imperfect, noncontiguous, 9-bp repeat, (CCAAAATCT)GC(CCCAAATCT), in the inferred ancestor of the RG population (labeled "RG anc" in Figure 1) and that this noncontiguous repeat provided the raw material for repeat expansion due to SSM. A T-to-C mutation in one motif of this noncontiguous repeat appears to have prevented repeat expansion in the ancestor of RG5 (Figure 1).

We have not sequenced the entire control region for RG5, the "nonexpanded" individual from the Rio Grande. Phylogenetic analysis of 150 bp of all 46 guppies sequenced (Table 2) places RG5 in a monophyletic group that includes Oropuche River, Quare River, and other Rio Grande samples. Based upon the observation that RG5 shares the A-to-C mutation described above with other RG samples, we have drawn a tree showing RG monophyly in Figure 1. Our hypothesis for the formation of the noncontiguous repeat does not depend upon our assumption that the RG clade is monophyletic.

Our model for expansion at this locus is presented in Figure 2. This model is similar to the SSM model (Levinson and Gutman 1987) and the illegitimate elongation model (Buroker *et al.* 1990) but emphasizes misalignment at noncontiguous repeats. First (Figure 2A), replication of the mitochondrial heavy strand (H-strand) pauses after termination associated sequences, producing a stable triple-stranded structure referred to as the D-loop (Shadel and Clayton 1997). Second (Figure 2B), competition between the H-strand and the D-loop strand for light strand (L-strand) binding facilitates local melting and the D-loop strand becomes single stranded. Local melting is followed by "competitive misalignment," *i.e.*, reinvasion of the D-loop strand and misannealing between the D-loop strand and L-strand (Buroker *et al.* 1990) (Figure 2C). Misannealing of the D-loop strand may be enhanced by hairpins that reduce the "effective" length of the D-loop strand, preventing it from reannealing with its proper L-strand complement (Buroker *et al.* 1990). After competitive misalignment, the continuation of D-loop strand elongation leads to the duplication of the intervening base pairs and one motif of the noncontiguous repeat producing a heteroduplex (Figure 2D). D-loop strand elongation, *i.e.*, formation of the nascent H-strand (H′ in Figure 2E), exposes the origin of replication for the L-strand (Shadel and Clayton 1997) leading to L-strand replication (Figure 2E). If the heteroduplex shown in Figure 2E is not repaired, then the next mtDNA replication produces three wild-type genomes (not shown) and a mitochondrial genome with a tandem repeat of an 11-bp motif flanked by one 9-bp motif from the original noncontiguous repeat (Figure 2F). We hypothesize that this is followed by expansion of the perfect tandem repeat via SSM. We refer to the

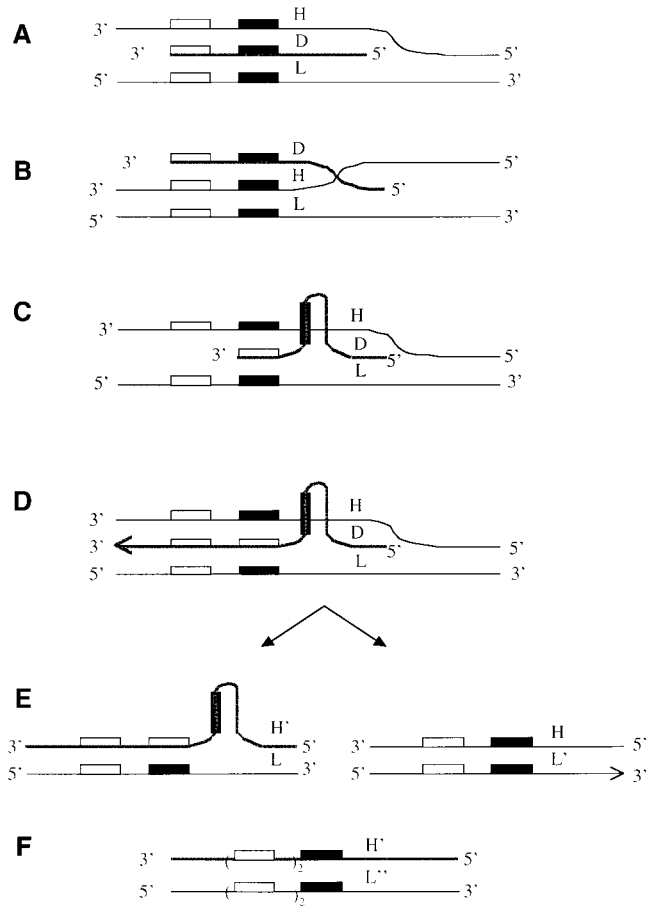

Figure 2.—Slipped-strand mispairing at a noncontiguous repeat in the mitochondrial control region. Noncontiguous repeat motifs are represented by shaded and unshaded rectangles. H, heavy strand; D, D-loop strand; L, light strand; H′, new H-strand formed by extension of D-loop strand; L′, light strand formed from H template in next round of replication; L″, light strand formed from the H′ template in the second round of replication. Arrows indicate the direction of replication. (A) The arrest of D-loop strand elongation leads to the formation of a stable triple-stranded structure called the D-loop. (B) Competitive misalignment (*i.e.*, displacement of the D-loop strand by the heavy strand). (C) Misalignment during reinvasion by the D-loop strand. (D) D-loop strand elongation after competitive misalignment duplicates the intervening nucleotides and one motif of the noncontiguous repeat, producing a heteroduplex. (E) Further extension of the D-loop strand produces a new H-strand (H′). (F) Replication of this H′-strand produces a mitochondrial genome containing a tandem repeat with a 3′ partial repeat on the light strand.

A 5′-(CCAAAATCTGC)$_{4-9}$
CCCAAATCTAT-3′

B 5′-(TTGCAAGTAT)$_{6}$
TTGCAAAGCA-3′

C 5′-(ATGCTATGTTTAATCCACATTAATTTCTAGCCACCATACCATAATGCTCACAAGCACATTAAATTGTTTAAGTACATAAGCA)$_{4}$
ATGCTATGTTTAATCCACATTAATTTCTAGCCATCATATCACAATGTTTCATCTACCATTAAATTGTTATACACCATTATTT-3′

5′-(ATACTATGTTTAATCCACATTAATCTCTAGTCACCATACCATAATGTTTGTAAATACATTAAATTACCTATATAGGAC)$_{2}$
ATACTATGTTTAATCCACATTAATCTCTATGTGCACTAACATACGATTTCCCGATACTTAGATATGTAGTAAGAGCCG-3′

5′-(ATATTATGTTTAATCCACATTAATTTCTAGTCATCATACATTAATGCTCGTACATACATTAAATTGTTTAAGTACATAGGAC)$_{3}$
ATATTATGTTTAATCCACATTAATTTCCAGTCACTACACACAGAATGTTTCATCTACCATCAAATGATCCACACCATTATCT-3′

D 5′-(CATATAAGCAAGTACTATTAATCTAATTAGTACATTAGACATAYTATGTATATCGTACATTAYATTCTYGWCCYCATR)$_{>4}$
CATATAAGCAAGTACATATAAAGTTTAATTATACATAATACATTAATTCCTTAATCGGACATAGCACATCTAGTGAAA-3′

E 5′-(TCTATTCCACATGCATATAAGCATGTACATACTACTATGATAGTACATAATACATTTTATGTATATCGTACATTAAGT)$_{4}$
TCTATTCCACATGCATATAAGCATGTACATACCATTATAACAGTACATAATACATTTTATTATTTATCGTACATAGGA-3′

F 5′-(TACATACTATGTATAATCAACATTCATACTATGYTTTTTA)$_{5}$
TACATACTATGTATAATCAACATTATGATAAATGAGGACT-3′

G 5′-(ATTCCAACCCTAAAATAACGGTGACCTACCCCAGACTTTCCAATTATTGTAATAATGTTAAT)$_{n}$
ATTCCAACCCTAAAATAACGGTGACCTACCCCCCCCCGAGCACTCCAACTATTAGCCTCTTCA-3′

H 5′-(ACTTTTTCAACCCTAAAATAGCGGTAGCCCACCCCAAGCTTACCGATTATTGTAATAATGTT)$_{n}$
ACTTTTTCAACCCTAAAATAGCGGTAGCCCACCCCCCCGGCAATTCAAGTATTGTTTTCTT-3′

Figure 3.—Mitochondrial DNA VNTRs and 3′ partial repeats. L-strand sequences shown. Partial repeats are underlined. Flanking sequences, equal in length to VNTR sequences, are typed below the repeats so that partial and imperfect repeats (see text) can be easily compared. Where nucleotides vary within a taxon the following code is used: Y, C or T; W, A or T; R, A or G. (A) A VNTR with an 11-bp motif in guppies (*P. reticulata*; sample RG1) and a 9-bp 3′ partial repeat (present study). (B) A VNTR with a 10-bp motif in zander (*Stizostedion lucioperca*) and a 6-bp 3′ partial repeat. Perch (*Perca fluviatilis*) and ruffe (*Acerina cernua*) also possess mtDNA VNTRs with 3′ partial repeats (Nesbø *et al.* 1998). (C) VNTRs in *Acipenser transmontanus*, *A. medirostris*, and *A. fulvescens* with 3′ partial repeats (Brown *et al.* 1996). (D) A VNTR with a 78-bp motif in *Crocidura russalu* (sample 101) and a 15-bp 3′ partial repeat. (E) A VNTR with a 78-bp motif in *Sorex araneus* (sample 4326) and a 32-bp 3′ partial repeat. The *C. russalu* VNTR includes the 3′ flanking repeat shown in Table 2 of Fumagalli *et al.* (1996) and the partial repeat (underlined) is not considered part of the locus by Fumagalli *et al.* (1996). (F) A VNTR with a 40-bp motif in the Taipei treefrog (*Rhacophorus taipeianus*; sample Y1) and a 24-bp 3′ partial repeat (Yang *et al.* 1994). (G) A VNTR with a 62-bp motif in yellowtail flounder (*Limanda ferruginea*) and a 32-bp partial repeat. (H) A VNTR with a 62-bp motif in winter flounder (*Pseudopleuronectes americanus*) and a 35-bp partial repeat (Lee *et al.* 1995). These flounder repeats occur in the R2 portion of the control region whereas all other mtDNA repeats in this figure are R1 repeats.

remnant of the noncontiguous repeat as a 3′ "partial" repeat (Figure 2F) because it occurs upstream of the repeat on the L-strand (L″ in Figure 2F).

DISCUSSION

The birth of an mtDNA minisatellite in guppies: The discovery of a highly variable mitochondrial minisatellite in one Trinidadian guppy population provided us with an opportunity to study the processes responsible for the formation of VNTRs. By mapping sequence data onto a guppy phylogeny we discovered mutations that appear to have produced a noncontiguous, imperfect repeat in the ancestor of guppies from the RG population. We propose that SSM at this noncontiguous repeat produced a tandem repeat with an 11-bp motif that was flanked by a 9-bp partial repeat, $(CCAAAATCTGC)_2CC$ CAAATCT, and that subsequent SSM at the tandem repeat led to the mtDNA length variation currently observed in the RG guppy population. Thus, the first SSM mutation produces a perfect tandem repeat with a motif that is longer than the noncontiguous repeat motif formed by nucleotide substitutions. This model is similar to a model proposed by Torroni *et al.* (1994) to explain the formation of a rare 207-bp duplication in the mitochondrial genome of Caucasian humans.

A general model for minisatellite birth? A survey of published minisatellite sequence data suggests that SSM at noncontiguous repeats may be a general model for VNTR birth in both mitochondrial and nuclear DNA. As described above, a consequence of SSM at a noncontiguous repeat is the formation of a locus with long repeats flanked by one unit of the original noncontiguous repeat (*i.e.*, a 3′ partial repeat). Fumagalli *et al.* (1996) noticed that in many taxa mtDNA VNTRs are flanked by "imperfect" or "degenerate" repeats. We compared the imperfect and "perfect" mtDNA repeats in shrews (Fumagalli *et al.* 1996), sturgeon (Brown *et*

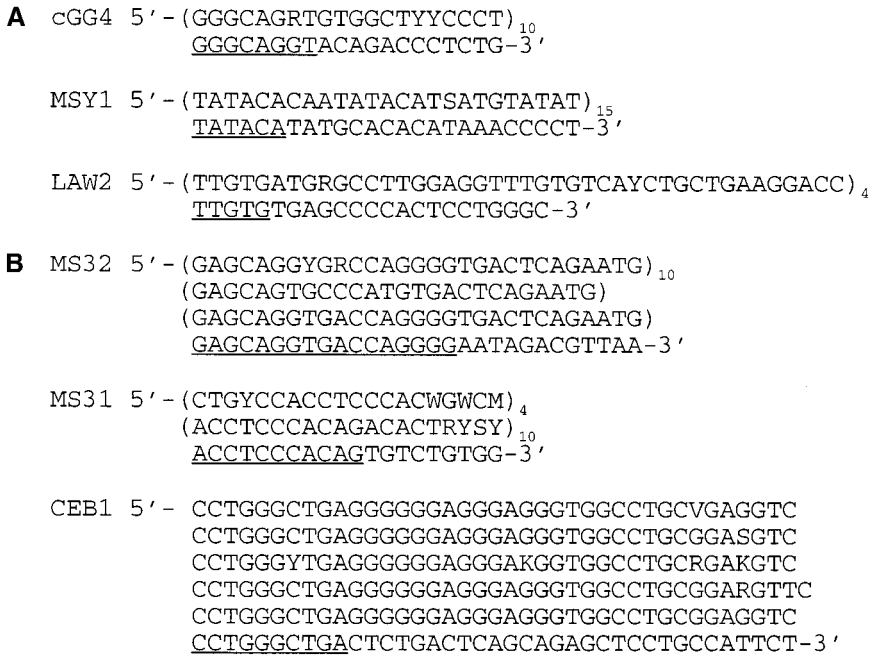

Figure 4.—Nuclear minisatellites and 3′ partial repeats in humans. Partial repeats are underlined. Where nucleotides vary within a taxon the following code is used: R, C or G; Y, C or T; S, C or G; W, A or T; M, A or C; K, G or T. (A) Sequences from Figure 2 in Haber and Louis (1998): cGG4 has a 20-bp motif repeated 10 times and an 8-bp 3′ partial repeat. MSY1 includes 15 25-bp motifs and is flanked by a 6-bp 3′ partial repeat. LAW2 includes four 42-bp repeats and a 5-bp partial repeat. (B) Sequence data described by Murray *et al.* (1999) and retrieved from GenBank (accession nos. AF048727–29): MS32 is composed of a 29-bp motif repeated 11 times, 1 26-bp motif, and a 17-bp 3′ partial repeat. MS31 is a compound minisatellite with one 20-bp motif repeated 4 times and a second 20-bp motif repeated 10 times. MS31 is flanked by an 11-bp 3′ partial repeat. CEB1 includes 4 39-bp motifs, 1 40-bp motif, and a 10-bp 3′ partial repeat.

al. 1996), perch (Nesbø *et al.* 1998), the Taipei treefrog (Yang *et al.* 1994), and flounder (Lee *et al.* 1995) and discovered that in all cases imperfect repeats may be interpreted as 3′ partial repeats. That is, the first portion of the imperfect repeat matches the perfect repeats best (Figure 3). Similarly, partial repeats (called "short direct" repeats by Haber and Louis 1998) occur next to nuclear minisatellites in humans (Haber and Louis 1998; Murray *et al.* 1999; Figure 4), birds (Gyllensten *et al.* 1989), salmon (Goodier and Davidson 1998), and fungi (Giraud *et al.* 1998). These observations suggest that the mtDNA-specific components of our model (*e.g.*, competitive misalignment) may not be critical and that the birth of VNTRs with long motifs in mitochondrial and nuclear DNA frequently involves SSM at noncontiguous repeats.

We thank Sampson Wu for technical assistance, Andrew T. Beckenbach and Michael J. Smith for valued discussions, and Ann E. Houde for sending additional Rio Grande guppies. This work was supported by a Simon Fraser University President's Ph.D. research stipend to J.S.T. and a Canadian National Sciences and Engineering Research Council grant to F.B.

LITERATURE CITED

Angers, B., and L. Bernatchez, 1997 Complex evolution of a salmonid microsatellite locus and its consequences in inferring allelic divergence from size information. Mol. Biol. Evol. **14:** 230–238.

Blanquer-Maumont, A., and B. Crouau-Roy, 1995 Polymorphism, monomorphism, and sequences in conserved microsatellites in primate species. J. Mol. Evol. **41:** 492–497.

Bowcock, A. M., A. Ruiz-Linares, J. Tomfohrde, E. Minch, J. R. Kidd *et al.*, 1994 High resolution of human evolutionary trees with polymorphic microsatellites. Nature **368:** 455–457.

Breden, F., M. B. Ptacek, M. Rashed, D. Taphorn and C. A. Figueiredo, 1999 Molecular phylogeny of a live-bearing fish genus *Poecilia* (Poeciliidae: Cyprinodontiformes). Mol. Phylogenet. Evol. **12:** 95–104.

Brohede, J., and H. Ellegren, 1999 Microsatellite evolution: polarity of substitutions within repeats and neutrality of flanking sequences. Proc. R. Soc. Lond. Ser. B **266:** 825–833.

Brown, J. R., K. Beckenbach, A. T. Beckenbach and M. J. Smith, 1996 Length variation, heteroplasmy and sequence divergence in the mitochondrial DNA of four species of sturgeon (*Acipenser*). Genetics **142:** 525–535.

Buroker, N. E., J. R. Brown, T. A. Gilbert, P. J. O'Hara, A. T. Beckenbach *et al.*, 1990 Length heteroplasmy of sturgeon mitochondrial DNA: an illegitimate elongation model. Genetics **124:** 157–163.

Dib, C., S. Fauré, C. Fizames, D. Samson, N. Drouot *et al.*, 1996 A comprehensive genetic map of the human genome based on 5,264 microsatellites. Nature **380:** 152–154.

Dietrich, W. F., J. Miller, R. Steen, M. A. Merchant, D. Damron-Boles *et al.*, 1996 A comprehensive genetic map of the mouse genome. Nature **389:** 149–152.

Fajen, A., and F. Breden, 1992 Mitochondrial DNA sequence variation among natural populations of the Trinidad guppy, *Poecilia reticulata*. Evolution **46:** 1457–1465.

Fumagalli, L., P. Taberlet, L. Favre and J. Hausser, 1996 Origin and evolution of homologous repeated sequences in the mitochondrial DNA control region of shrews. Mol. Biol. Evol. **13:** 31–46.

Giraud, T., D. Fortini, C. Levis and Y. Brygoo, 1998 The minisatellite MSB1, in the fungus *Botrytis cinerea*, probably mutates by slippage. Mol. Biol. Evol. **15:** 1524–1531.

Goodier, J. L., and W. S. Davidson, 1998 Characterization of novel minisatellite repeat loci in Atlantic salmon (*Salmo salar*) and their phylogenetic distribution. J. Mol. Evol. **46:** 245–255.

Gyllensten, U. B., S. Jakobsson, H. Temrin and A. Wilson, 1989 Nucleotide sequence and genomic organization of bird minisatellites. Nucleic Acids Res. **17:** 2203–2214.

Haber, J. E., and E. J. Louis, 1998 Minisatellite origins in yeast and humans. Genomics **48:** 132–135.

Hagelberg, E., I. C. Gray and A. J. Jeffreys, 1991 Identification of the skeletal remains of a murder victim by DNA analysis. Nature **352:** 427–429.

Higgins, D. G., A. J. Bleasby and R. Fuchs, 1992 CLUSTAL: a package for performing multiple sequence alignment on a microcomputer. Comput. Appl. Biosci. **8:** 189–191.

Kocher, T. D., W. K. Thomas, A. Meyer, C. V. Edwards, S. Paabo *et al.*, 1989 Dynamics of mitochondrial DNA evolution in animals: amplification and sequencing with conserved primers. Proc. Natl. Acad. Sci. USA **86:** 6196–6200.

Lee, W.-J., J. Conroy, W. H. Howell and T. D. Kocher, 1995 Struc-

ture and evolution of Teleost mitochondrial control regions. J. Mol. Evol. **41:** 54–66.

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

Lunt, D. H., L. E. Whipple and B. C. Hyman, 1998 Mitochondrial DNA variable number tandem repeats (VNTRs): utility and problems in molecular ecology. Mol. Ecol. **7:** 1441–1455.

Mandel, J.-L., 1997 Breaking the rule of three. Nature **386:** 767–769.

Messier, W., S. Li and C. Stewart, 1996 The birth of microsatellites. Nature **381:** 483.

Meyer, A., J. M. Morrissey and M. Schartl, 1994 Recurrent origin of a sexually selected trait in *Xiphophorous* fishes inferred from a molecular phylogeny. Nature **368:** 539–542.

Murray, J., J. Buard, D. L. Neil, E. Yeramian, K. Tamaki *et al.*, 1999 Comparative sequence analysis of human minisatellites showing meiotic repeat instability. Genome Res. **9:** 130–136.

Nesbø, C. L., M. O. Arab and K. S. Kakobsen, 1998 Heteroplasmy, length and sequence variation in the mtDNA control region of three percid fish species (*Perca fluviatilis*, *Acerina cernua*, *Stizostedion lucioperca*). Genetics **148:** 1907–1919.

Olaisen, B., M. Stenersen and B. Mevåg, 1997 Identification by DNA analysis of the victims of the August 1996 Spitsbergen civil aircraft disaster. Nat. Genet. **15:** 402–405.

Ptacek, M. B., and F. Breden, 1998 Phylogenetic relationships among the mollies (Poeciliidae: *Poecilia*: *Mollienesia* group) based on mitochondrial DNA sequences. J. Fish Biol. **53:** 64–81.

Schlötterer, C., C. Vogl and D. Tautz, 1997 Polymorphism and locus-specific effects on polymorphism at microsatellite loci in natural *Drosophila melanogaster* populations. Genetics **146:** 309–320.

Shadel, G. S., and D. A. Clayton, 1997 Mitochondrial DNA maintenance in vertebrates. Annu. Rev. Biochem. **66:** 409–435.

Swofford, D. L., 1993 PAUP: phylogenetic analysis using parsimony. Version 3.11 Illinois Natural History Survey, Champaign.

Taylor, J. S., J. M. H. Durkin and F. Breden, 1999 The death of a microsatellite: a phylogenetic perspective on microsatellite interruptions. Mol. Biol. Evol. **16:** 567–572.

Torroni, A., M. T. Lott, M. F. Cabell, Y.-S. Chen, L. Lavergne *et al.*, 1994 mtDNA and the origin of caucasians: identification of ancient caucasian-specific haplogroups, one of which is prone to a recurrent somatic duplication in the D-loop region. Am. J. Hum. Genet. **55:** 760–776.

Wada, C., S. Shionoya, Y. Fujino, H. Tokuhiro, T. Akahoshi *et al.*, 1994 Genomic instability of microsatellite repeats and its association with the evolution of chronic myelogenous leukemia. Blood **83:** 3449–3456.

Yang, Y.-J., Y.-S. Lin, J.-L. Wu and C.-F. Hui, 1994 Variation in mitochondrial DNA and population structure of the Taipei treefrog *Rhacophorus taipeianus* in Taiwan. Mol. Ecol. **3:** 219–228.

Zardoya, R. D., M. Vollmer, C. Craddock, J. T. Streelman, C. Karl *et al.*, 1996 Evolutionary conservation of microsatellite flanking regions and their use in resolving the phylogeny of cichlid fishes (Pisces: Perciformes). Proc. R. Soc. Lond. Ser. B **263:** 1589–1598.

Communicating editor: S. Yokoyama