11 | mRNA editing

LARRY SIMPSON and OTAVIO H. THIEMANN

1. Introduction

RNA editing is a term that was coined by Benne et al. (1) to describe the apparently post-transcriptional insertion of four Us at three sites in the coding region of the cytochrome oxidase subunit II (COII) mRNA in two kinetoplastid species, Crithidia fasciculata and Trypanosoma brucei. This modification overcame a −1 frameshift in the COII gene that had been shown to be evolutionarily conserved in these two species and also in Leishmania tarentolae (2). As additional examples of RNA sequence modifications were reported in a variety of organisms (2–10), RNA editing came to be defined in a broad sense as any process that results in the production of an RNA molecule which differs in nucleotide sequence in coding regions from the DNA template, with the exception of classical cis-splicing (11). However, even this definition was rendered obsolete by the discovery of specific sequence changes in rRNAs and tRNAs in several organisms, in one case involving the insertion of cytidine residues (12, 13) and in some cases the substitution of uridine for cytidine (14) or guanosine for adenosine residues (15). In this review we will only discuss the trypanosome type of RNA editing, which has been termed insertion–deletion editing by Bass (16), to distinguish it from the substitution editing found in other cells.

The kinetoplastid protozoa, together with the euglenoids, represent one of the most ancient eukaryotic lineages in terms of rRNA phylogenetic reconstructions (17), and possibly represent one of the surviving initial lineages which possessed a mitochondrion. The kinetoplasts consist of two major branches, the Bodonina (bodonids and cryptobiids) and the Trypanosomatina (trypanosomatids), both of which contain a single mitochondrion with a large mass of DNA. The region of the mitochondrion where the DNA is localized is termed the kinetoplast as a consequence of its association with the flagellum, and its DNA is called kinetoplast DNA (kDNA).

Little was known until recently about the kDNA of the bodonids. The kDNA of Bodo caudatus, a free-living bodonid flagellate, was shown in 1986 (18) to consist of heterogeneous sized large molecules of unknown function. More recently, the kDNA of a parasitic cryptobiid flagellate, Trypanoplasma borreli, was isolated and partially characterized (19, 20). The kDNA of T. borreli consists of a 40–80 kb circular component that contains several structural genes and several larger circular components consisting of tandemly organized 1 kb sequences encoding small RNAs which may be involved in the mechanism of RNA editing (see Section 3).
The kDNA of the trypanosomatids has been studied intensively using species from several genera (21). It consists of thousands of catenated small circular molecules, the minicircles, and a smaller number of catenated larger circular molecules, the maxicircles. The minicircle size varies in different species from approximately 0.5 kb to 2.5 kb and the maxicircle varies from 22 kb to 36 kb. The maxicircle molecule is the homologue of the mitochondrial DNA found in animal cells and the minicircle molecule encodes the guide RNAs (gRNAs) involved in mediating editing of the transcripts of certain maxicircle structural genes (22). These genes whose transcripts are edited are termed ‘cryptogenes’ (11).

2. Mitochondrial cryptogenes in kinetoplastids

The maxicircle genome is organized into an informational region that contains the two rRNA genes (23–26), at least 11 structural genes, and an apparently non-informational region that contains tandem repeats of varying complexities (21, 27–35) (Fig. 1). The latter region differs in size and sequence in different species and accordingly is termed the ‘divergent’ or ‘variable’ region. Several maxicircle genes yield transcripts which are apparently never edited: cytochrome oxidase subunit I (COI); NADH dehydrogenase subunits 1 (ND1), 4 (ND4), and 5 (ND5); and maxicircle unidentified reading frames 1 (MURF1) and 5 (MURF5). MURF5 (previously called ORF10) (31) in L. tarentolae is an apparently unedited 100 amino acid open reading frame situated between G2 and ND7, in the same polarity as G2. It exhibits a codon bias characteristic of a coding region and has significant similarity to the translated equivalent open reading frame in T. brucei (previously called ORF8) (31), but has no homology with any database sequence. The HR2 nucleotide sequence, which represents a sequence fairly well conserved between L. tarentolae and T. brucei, lies within this region (31).

Cryptogenes produce transcripts which are edited, and the extent of editing varies from gene to gene. Those cryptogenes, transcripts of which are only slightly or moderately edited in L. tarentolae and C. fasciculata, include cytochrome b (Cyb) (4, 5, 36), cytochrome oxidase subunits II (COII) (1, 2) and III (COIII), ND7, MURF2, and MURF4 (3, 9). In T. brucei, the ND7, COIII, and MURF4 genes are not recognizable at the genomic sequence level (31), but the transcripts are extensively edited or ‘pan-edited’ (6–8) (Section 2.1) to produce potentially translatable sequences which have homology with known genes. In addition to these cryptogenes, there are six G-rich intergenic regions (labelled G1–G6 in L. tarentolae and CR1–CR6 in T. brucei) which in T. brucei and in a recently isolated strain of L. tarentolae (LEM125) yield transcripts that are pan-edited (11, 37–40). The G6 transcript is edited to produce an mRNA encoding the ribosomal protein, S12; the edited G1, G2, and G5 transcripts encode components of complex I of the electron transport chain—ND8, ND9, and ND3, respectively; and the edited G3 and G4 transcripts encode hydrophobic proteins of unknown function.
2.1 Extent of editing and identification of gene products of edited mRNAs

The extent of editing of cryptogene transcripts ranges from the insertion of four or five Us in three sites in the middle of the gene (internal editing), to the addition of 30–40 Us at 15–20 sites at the 5' end of the gene (5' editing), to pan-editing (11) over the entire length of the gene. Initially it seemed that pan-editing was limited to the ND7, COIII, and MURF4 cryptogenes in T. brucei. However, the 5' third of the MURF4 mRNA (6, 41) has been found to be pan-edited in L. tarentolae, in addition to all the G1–G6 mRNAs (40).

2.1.1 The MURF4 cryptogene

In T. brucei, the 821 nucleotide MURF4 pre-edited mRNA is pan-edited with the addition of 448 Us at 173 sites and the deletion of 28 Us at 12 sites (6, 41). Editing occurs both within the coding region and within the 3' untranslated region, and terminates one nucleotide upstream of the created AuG methionine translation initiation codon. The 5' end (32 nucleotides) of the untranslated 5' flanking sequence
and the 3' terminal portion (37 nucleotides) of the 3' untranslated sequence remain unedited. It is not known if there are separate editing domains, which are defined as regions of the RNA which are edited independently of other regions, since an analysis of multiple partially edited MURF4 transcripts has not yet been performed.

The editing in the case of the L. tarentolae MURF4 transcript is less extensive, but still substantial (6, 41). Pan-editing is limited to the 5' portion of the mRNA, and involves the addition of 106 Us at 46 sites and the deletion of 5 Us at four sites within a 112 nucleotide region. An AuG methionine initiation codon is created at the same relative position as in T. brucei.

The pan-edited RNA from T. brucei contains an ORF which shows high similarity with the translated protein from the 5' pan-edited L. tarentolae RNA. The originally reported T. brucei MURF4 sequence (6) indicated a substantial divergence of the translated amino acid sequence at the C terminus of the protein. However, a revised T. brucei sequence (41) corrected for a frameshift error showed that the similarity in translated amino acid sequences continues to the C terminus, with 65% identity and 85% overall similarity, allowing for conservative replacements. From a multiple sequence alignment with known ATPase 6 proteins, Bhat et al. (6, 41) suggested that the MURF4 protein belongs to this family of proteins, although highly diverged. However, the similarity of the kinetoplastid MURF4 amino acid sequences with known ATPase 6 sequences is of marginal statistical significance if one applies a Monte Carlo shuffling analysis (42, 43) (L. Simpson, unpublished). On the other hand, a PROFILESEARCH analysis (44) of the Swiss Protein database gave 12 known ATPase 6 proteins as the best 'hits' with Z values of 5.0–7.7 standard deviation (sd) units from the mean of aligned random sequences (M. Peris, and L. Simpson, unpublished). It is entirely possible that MURF4 represents a highly diverged ATPase 6, but this cannot be confirmed by sequence comparisons alone and rather must be substantiated by sequence analysis of the kinetoplastid ATPase 6 protein.

2.1.2 The G6 (CR6) cryptogene

Another pan-edited cryptogene in L. tarentolae is the G6 region. Editing occurs in three separate 'editing domains' (Fig. 2) in both UC strain and LEM125 strain L. tarentolae. The mature edited L. tarentolae G6 mRNA contains 117 Us inserted at 49 sites and 32 Us deleted at 13 sites (39). The pan-edited mRNA encodes an 85 amino acid polypeptide which appears to represent a highly diverged ribosomal protein S12 (RPS12). In T. brucei, 132 Us are inserted and 28 Us are deleted, yielding a predicted 82 amino acid protein with good similarity to the L. tarentolae G6 protein sequence (38). The two most conserved regions of the protein sequences correspond to the sites which are known to confer streptomycin resistance and streptomycin dependence. Marginally significant Z values of 6.5–8.9 sd units were obtained for alignments of the G6 sequence with three chloroplast S12 proteins and one eubacterial S12 protein. Statistically insignificant Z values, however, were obtained for the alignments of G6 with mitochondrial S12 proteins from Paramecium and Zea mays. The conclusion that the G6 protein belongs to the RPS12 family is strengthened by
Fig. 2 The pan-edited RPS12 (G6) cryptogene from L. tarentolae. The RPS12 DNA sequence is nucleotides 14,636 to 14,913 in GenBank entry LEIKPMAK. Adjacent portions and polarities of the G5 and ND5 genes are shown by arrows. The sequence of mature edited RPS12 mRNA is shown, with editing domains I, II, and III indicated by stippling. Uridines added by editing are shown as 'u'. The amino acid sequence of the translated RPS12 protein is beneath the edited mRNA sequence. The domain-connection sequences DCS-I and DCS-II are boxed. (Reprinted with permission from 39.)
similarities in hydropathy patterns, but as in the case of MURF4, the amino acid sequence of the RPS12 protein is required to confirm this assignment.

The presence of three editing domains in the RPS12 cryptogene in *L. tarentolae* opens the possibility that independent editing of separate domains may modulate amino acid sequences of proteins in addition to simply creating translatable mRNAs. Another example of a cryptogene with more than one editing domain is the ND7 pan-edited cryptogene of *T. brucei* (7). In *L. tarentolae* and *C. fasciculata*, this gene has an internal frameshift editing site and a 5’ editing site, which also may be considered two separate editing domains, especially since these sites occur at the equivalent of the 5’ termini of the two pan-edited domains in *T. brucei* (2, 3) (Fig. 3).

Alignment of the mature edited RPS12 RNA sequences from *L. tarentolae* and *T. brucei* shows that the two domain-connection sequences (DCSs), which may represent the functionally significant regions of the protein, show absolute sequence conservation at the amino acid level (Fig. 4). However, at the RNA level, the DCSs are edited in *T. brucei*, thereby eliminating the possibility that the existence of separate editing domains in *L. tarentolae* is due to a functional constraint on editing in the DCS regions. This lack of correspondence of editing domains between these two species also makes it less likely that, at least in this case, independent regulation of editing in separate domains is used as a regulatory mechanism. Unedited DCSs were also not present in the pan-edited RPS12 sequence of the cryptobiid, *T. borreli* (19).

### 2.1.3 The G1 (CR1) cryptogene

The CR1 region of the maxicircle genome of *T. brucei* produces a 361 nucleotide pre-edited transcript that is pan-edited by the addition of 259 Us and the deletion of 46 Us at a total of 127 sites (37). The 5’ terminal 34 nucleotides and the 3’ terminal 27 nucleotides remain unedited, as is the case with all other reported pan-edited mRNAs. Editing continues four sites upstream of the created AuG initiation codon. The open reading frame encodes a 145 amino acid protein that shows significant similarity to a nuclear-encoded subunit of the bovine mitochondrial respiratory complex I, CI-23kD, including a repeated cysteine motif characteristic of a class of non-haem iron sulphur proteins. From this homology, the *T. brucei* protein was identified as NADH dehydrogenase subunit 8 (ND8).

Thiemann et al. (40) showed that productive editing of G1 transcripts also occurs in the LEM125 strain of *L. tarentolae*. The mature edited G1 RNA of LEM125 is 520 nucleotides in length, and is pan-edited by the addition of 215 Us in 99 sites and the deletion of 41 Us in 17 sites. An open reading frame (ORF) of 145 amino acid residues which is encoded by the edited transcript is homologous to the ND8 polypeptide encoded by the edited G1 (CR1) RNA in *T. brucei* (37).

### 2.1.4 The G2 (CR2) cryptogene

A 322 nucleotide transcript of the CR2 cryptogene of *T. brucei* is edited by the addition of 345 Us and the deletion of 20 encoded Us, yielding a mature edited mRNA encoding a 194 amino acid protein. This protein has significant similarity to a sub-
Fig. 3 Comparison of editing profiles in mature edited ND7 RNA sequences from *L. tarentolae* (Lt.) and *T. brucei* (T.b.). Matches are shown by vertical lines and dashes indicate the gaps introduced for alignment. The translated amino acid sequences are shown above and below the RNA sequences. Edited regions are stippled, and the DCS sequences are boxed. Only the 5' portion of domain of the *T. brucei* sequence is shown. (Reprinted with permission from 111.)
Fig. 4 Alignments of RPS12 sequences from *L. tarentolae* (L.t.) and *T. brucei* (T.b.) Edited regions are stippled, inserted uridines are indicated as 'u', and deleted uridines as '*'. The DCS sequences in *L. tarentolae* are boxed. The amino acid sequences of the translated proteins are shown above and below the edited RNA sequences. (Reprinted with permission from 111.)
unit of complex I from other organisms and was termed NADH dehydrogenase subunit 9 (ND9) (45). The transcript of the G2 cryptogene of L. tarentolae LEM125 (40) contains 335 U additions in 125 sites and 40 deletions in 15 sites, producing a mature edited mRNA encoding a 196 amino acid sequence. This sequence shows significant similarity to the ND9 sequence from T. brucei.

2.1.5 The G3 (CR3) cryptogene
The G3 genomic sequence is the shortest cryptogene in the maxicircle and, at least in the case of L. tarentolae LEM125 (40), shows some unusual features. The transcript is edited by 35 U additions and 14 U deletions. The deletions occur in the last three sites and the intervening genomic sequence is G-rich, suggesting incomplete editing. The edited RNA encodes an ORF of 51 hydrophobic amino acids which has no homology to any database sequence.

2.1.6 The G4 (CR4) cryptogene
A 283 nucleotide transcript of the CR4 cryptogene of T. brucei is edited by the insertion of 352 Us and the deletion of 40 Us to yield an edited mRNA encoding two possible ORFs with no detectable homology to any database sequence. In the case of L. tarentolae LEM125, the G4 transcript is edited by the addition of 326 Us in 110 sites and the deletion of 5 Us in two sites. A predicted ORF is 169 amino acids long and exhibits a limited similarity with one of the ORFs from the edited T. brucei mRNA.

2.1.7 The G5 (CR5) cryptogene
The transcript of the CR5 cryptogene of T. brucei is edited in two separate domains separated by an 8 nucleotide unedited DCS (46). The 3' domain shows several different editing patterns whereas the 5' domain has a single consensus pattern. A total of 205-217 Us are inserted and 13-16 Us are deleted in both domains. The existence of several different editing patterns in domain I would give rise to multiple carboxy-terminal protein sequences if all were translated, and is reminiscent of a misediting situation. Misediting is the occurrence of unexpected editing patterns which differ from the mature edited sequence (47). The edited transcript encodes a potential protein that shows a motif characteristic of iron-sulphur proteins which are present in complex I of the respiratory chain, and exhibits a limited similarity to ND3 sequences from other organisms.

In L. tarentolae LEM125, there appears to be a single editing domain in G5, in which 167 Us are inserted at 71 sites and 5 Us are deleted at three sites. The mature edited mRNA encodes an ORF of 115 amino acids which has significant similarity to the T. brucei ND3 sequence.

2.2 Addition of Us to 3’ ends of mRNAs
Many maxicircle transcripts occur as two distinct size classes (4, 5, 36, 48-51). This size variation has been shown to be due mainly to variation in the length of the poly(A) tail (41, 52). The transcripts include CYb, COI, and COII from T. brucei,
MURF4, CYb, and COIII from *L. tarentolae*, and ND1, MURF2, ND7, and COII from *C. fasciculata*. The two MURF4 transcripts of *L. tarentolae* contain 20 nucleotide and 120 nucleotide poly(A) tails (41, 52). In addition, variable numbers of U residues are found within the poly(A) tails (53). Since the patterns of U insertion differ in different cDNAs and several cDNAs show 3' terminal Us, it is likely that this U addition is not gRNA-mediated, but is due to a 3' terminal uridylyl transferase (TUTase) activity which is present in the kinetoplast-mitochondrion (54).

In the initial report of the mitochondrial TUTase activity (54), it was shown that the major endogenous RNA species labeled when isolated organelles are incubated with [$\alpha^{32}$P]UTP are the 9S and 12S mitochondrial rRNAs (54). This observation was confirmed and extended for *T. brucei* by Adler et al. (55) who showed that the 12S RNA has a 3' tail of 2–17 Us whereas the 9S RNA has a 3' tail of precisely 11 Us. This suggests the possibility of a more precise mechanism for 3' U addition in this case. It was speculated that the functional significance of this 3' U addition to the rRNAs may reside in modifications of the secondary or tertiary structures of the rRNAs, in protection against exonuclease degradation, or may even involve interactions with mRNAs (55).

2.3 Developmental regulation of RNA editing in the life cycle of *T. brucei*

The life cycle of the African trypanosome, *T. brucei*, involves several developmental stages which differ dramatically in terms of mitochondrial physiology, among other features (56). The procyclic trypomastigote forms found in the tsetse fly midgut possess an active mitochondrion that has a functional cytochrome-based respiratory chain, whereas the slender bloodstream trypomastigote forms found in the mammalian bloodstream have an inactive mitochondrion that lacks detectable cytochrome oxidase. Respiration of the latter forms is through a cyanide-insensitive alternate oxidase, and oxidative phosphorylation is absent (57). Several of the mitochondrial cryptogenes—MURF4, MURF2, G4 (CR4), and COIII—are constitutively edited in both life cycle stages, whereas CYb and COII are edited only in the procyclic stage (4, 6–8, 50, 58). A third set of cryptogene transcripts—RPS12, ND8, ND9, and CR3—are fully edited only in the bloodstream stages (37, 38, 59). The ND7 transcript shows a more complex pattern of regulation, in that domain I is fully edited only in the bloodstream stages and domain II is edited in both stages (7). The level of regulation of editing is uncertain. In several cases, the relative gRNA abundances for regulated transcripts do not change between life-cycle stages, suggesting that regulation is not at the level of gRNA abundance (58). However, in the only cases of pan-edited genes in which the entire gRNA cascade is known (the RPS12 and MURF4 genes of *L. tarentolae*) a single late-acting gRNA is in low abundance relative to the other gRNAs, opening the possibility that pan-editing might be regulated by the relative abundance of a single gRNA in the cascade (39).
Developmental regulation of the extent of 3' polyadenylation of maxicircle transcripts has also been demonstrated in T. brucei. A larger fraction of CYb, COI, and COII transcripts have longer poly(A) tails and are more abundant in procyclic than in bloodstream forms (52). On the other hand, a greater number of CR1 transcripts have longer poly(A) tails and are more abundant in bloodstream than in procyclic forms. Edited transcripts have longer poly(A) tails than unedited transcripts. The ND4 and MURF1 transcripts have a similar size distribution of poly(A) tails in both stages. Bhat et al. (52) speculated that regulation of polyadenylation may influence mitochondrial gene expression in this species.

Developmental regulation of editing has not been studied in the case of L. tarentolae due to the inability to cultivate the lizard stages of the life cycle, and to a lack of knowledge of the biology of the complete life cycle in the lizard host. In the case of monogenetic species such as C. fasciculata, which lack a vertebrate host cycle, it would be expected that editing is unregulated. However, there could well be physiological changes of the parasite within the insect host, which involve regulation of editing at some level. This remains to be investigated.

3. Guide RNAs

Guide RNAs (gRNAs) are small mitochondrial RNAs which can form perfect hybrids with mature edited mRNAs if G-U (and, in some cases, A-C) base pairs are allowed (22) (Figs 5 and 6). The region of base complementarily extends 3' of the pre-edited region for a variable length, and the formation of the 3' anchor RNA-RNA duplex was hypothesized to represent the initial interaction between a specific gRNA and a specific pre-edited mRNA. The gRNAs also possess non-encoded 3' oligo(U) tails, 5-30 nucleotides in length, and can be recognized on acrylamide-urea gels by a characteristic multiple banding pattern migrating ahead of tRNAs (60). They also have 5'-di- or tri-phosphates and can be capped in vitro by vaccinia virus guanylyl transferase.

Minicircle-encoded gRNAs were actually first visualized on northern blots as contaminants of L. tarentolae mitochondrial tRNA preparations, but the significance of this observation was not appreciated at the time (61).

3.1 Maxicircle-encoded gRNAs

The existence of seven maxicircle-encoded gRNAs was initially predicted by a computer analysis of the known 21 kb L. tarentolae maxicircle sequence for short sequences that could give rise to transcripts that were complementary to mature edited mRNA sequences if G-U base-pairing was allowed (Table 1). This was confirmed by northern blot hybridization using synthetic oligomer probes and by direct 5' end sequencing of hybrid-selected isolated gRNAs (22). The locations of the gRNA genes in the maxicircle genome have no relationship to the location of the homologous cryptogenes. Isolated gMURF2-II was shown to possess a di- or tri-phosphate at the 5' end and therefore may represent a primary transcript (60). All
pND8-II (mc)

3'...C----UagCaaCaagaUacgaU---AAACUaagaAUAAACAAA--5'

pND8-I(mc)

3'...guagelUaCagaUacgaUCagUcaCaCaG---AGCUCAUCAAINA--5'

DNA 5'... GTTTG G G A G ATTTTTTA A G G G A G A CTTTGAGAATATTTGATTATTAAATTAAATTAAATTAAATT 3'

RNA 5'... uuG****GuuGuuGuuAuGuuA**UUGAuuuuuAuuuGuuGuuGuuAuAGuuAGuAuC**UCGAAGAUAUUUGAUAAUUUAUUUAUUAUUUUU(A). 3'

Fig. 5 Two overlapping gRNAs mediating the editing of blocks I and II of ND8 mRNA. The duplex anchor sequences are shaded. G-U base-pairs are indicated by: and A-U and G-C base-pairs by |. Editing sites are numbered 3' to 5'. Deletions are indicated by *. The 3'-oligo(U) tails of the gRNAs are not shown.
maxicircle-encoded gRNAs in *L. tarentolae* appear to terminate, or to be 3\'-processed, at a string of encoded U residues, but a low percentage of 3\'-truncated gRNAs were also found in steady-state kRNA. An additional maxicircle-encoded gRNA (M150) of unknown function was detected in the UC strain of *L. tarentolae* due to its presence in a misedited COIII mRNA–gRNA chimeric molecule (47) (see below). Five additional *L. tarentolae* maxicircle-encoded gRNAs—gND9-XIV, gG3-I, gG3-II, gG4-IV, and gND3-I—were recently identified by computer analysis (40). These gRNAs were confirmed by northern analysis and primer-extension sequencing (40).

The gRNA for COII was found to be in cis at the 3\' terminus of the mRNA, and a fold-back mechanism for mediation of editing was proposed in this case (22).

Seven gRNA genes located at identical relative positions in the maxicircle...
### Table 1: Identified gRNAs from *L. tarentolae* UC and LEM125 strains

<table>
<thead>
<tr>
<th>mRNA</th>
<th>gRNA from L. tarentolae UC</th>
<th>gRNA from L. tarentolae LEM125</th>
</tr>
</thead>
<tbody>
<tr>
<td>RPS12 (G6)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>gRPS12-I</td>
<td>mc&lt;sup&gt;b&lt;/sup&gt;</td>
<td>gND8-I</td>
</tr>
<tr>
<td>gRPS12-II</td>
<td>mc</td>
<td>gND8-II</td>
</tr>
<tr>
<td>gRPS12-III</td>
<td>mc</td>
<td>gND8-III</td>
</tr>
<tr>
<td>gRPS12-IV</td>
<td>mc</td>
<td>gND8-IV</td>
</tr>
<tr>
<td>gRPS12-V</td>
<td>mc</td>
<td>gND8-V</td>
</tr>
<tr>
<td>gRPS12-VI</td>
<td>Mc (17020–16975)&lt;sup&gt;c&lt;/sup&gt;</td>
<td>gND8-VII</td>
</tr>
<tr>
<td>gRPS12-VII</td>
<td>mc</td>
<td>gND8-IX</td>
</tr>
<tr>
<td>gRPS12-VIII</td>
<td>mc</td>
<td>gND8-X</td>
</tr>
<tr>
<td>MURF2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>gMURF2-I</td>
<td>Mc (9908–9893)</td>
<td>gND8-XII</td>
</tr>
<tr>
<td>gMURF2-II</td>
<td>Mc (13087–13146)</td>
<td>gND8-XIII&lt;sup&gt;a&lt;/sup&gt;</td>
</tr>
<tr>
<td>MURF4 (ATPase 6)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>gMURF4-I</td>
<td>mc</td>
<td>gND9-II</td>
</tr>
<tr>
<td>gMURF4-II</td>
<td>mc</td>
<td>gND9-III</td>
</tr>
<tr>
<td>gMURF4-III</td>
<td>mc</td>
<td>gND9-V</td>
</tr>
<tr>
<td>gMURF4-IV</td>
<td>mc</td>
<td>gND9-VI</td>
</tr>
<tr>
<td>gMURF4-V</td>
<td>mc</td>
<td>gND9-VII</td>
</tr>
<tr>
<td>gMURF4-VI</td>
<td>mc</td>
<td>gND9-VIII</td>
</tr>
<tr>
<td>COII</td>
<td></td>
<td></td>
</tr>
<tr>
<td>gCOII-I</td>
<td>Mc (10120–10148)</td>
<td>G3</td>
</tr>
<tr>
<td>gCOII-II</td>
<td>mc</td>
<td>gG3-I</td>
</tr>
<tr>
<td>gCOII-III</td>
<td>mc</td>
<td>gG3-II&lt;sup&gt;a&lt;/sup&gt;</td>
</tr>
<tr>
<td>gCOII-IV</td>
<td>mc</td>
<td>gG3-III</td>
</tr>
<tr>
<td>CYb</td>
<td></td>
<td></td>
</tr>
<tr>
<td>gCYb-I</td>
<td>Mc (16803–16767)</td>
<td>G4</td>
</tr>
<tr>
<td>gCYb-II</td>
<td>Mc (2290–2239)</td>
<td>gG4-I</td>
</tr>
<tr>
<td>gCYb-III</td>
<td>mc</td>
<td>gG4-II</td>
</tr>
<tr>
<td>gCYb-IV</td>
<td>mc</td>
<td>gG4-III</td>
</tr>
<tr>
<td>ND3 (G5)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>gND3-I</td>
<td>Mc (304–350)</td>
<td>G4 IV</td>
</tr>
<tr>
<td>gND3-II</td>
<td>mc</td>
<td>gG4-V</td>
</tr>
<tr>
<td>gND3-III</td>
<td>mc</td>
<td>gG4-VI</td>
</tr>
<tr>
<td>gND3-IV</td>
<td>mc</td>
<td>gG4-VII</td>
</tr>
<tr>
<td>gND3-V</td>
<td>mc</td>
<td>gG4-VIII</td>
</tr>
<tr>
<td>gND3-VI</td>
<td>mc</td>
<td>gG4-IX</td>
</tr>
<tr>
<td>gND3-VII</td>
<td>mc</td>
<td>gG4-X</td>
</tr>
<tr>
<td>gND3-VIII</td>
<td>mc</td>
<td>gG4-XIV</td>
</tr>
<tr>
<td>ND7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>gND7-I</td>
<td>Mc (16724–16752)</td>
<td>Unassigned</td>
</tr>
<tr>
<td>gND7-II</td>
<td>Mc (395–346)</td>
<td>gM150&lt;sup&gt;f&lt;/sup&gt;</td>
</tr>
</tbody>
</table>

<sup>a</sup> The cryptogenes are indicated in bold.

<sup>b</sup> mc: minicircle-encoded gRNA.

<sup>c</sup> Mc: maxicircle-encoded gRNA. The position of the gene in the *L. tarentolae* maxicircle (LEIKPMAX) sequence is indicated in parenthesis.

<sup>d</sup> Putative maxicircle-encoded gRNA for NDS (G1) Block XIII.

<sup>e</sup> Putative maxicircle-encoded gRNA for G3 Block II.

<sup>f</sup> Putative gRNA found in a gRNA–mRNA misguided chimera.
genome were identified by sequence analysis alone for *C. fasciculata* (62). The presence of several compensatory base changes in anchor regions that preserve base-pairing with mRNA transcripts from the same species provides strong evidence for the existence of these gRNAs and for the importance of the anchor region for the editing process itself (62). In the *T. brucei* maxicircle sequence, only the gMURF2-I and II and gCOII genes are present, with the gND7-I and II, and the gCYb-I and II genes being missing (62).

A computer algorithm was developed based on the pairwise local similarity algorithm of Smith and Waterman for finding gRNAs given the cryptogene sequence (63). A test of this method with four known cryptogenes from *L. tarentolae* showed that additional information was required for accurate identification of gRNA sequences. The statistical distribution of the longest candidate gRNA sequences showed that the average expected length of gRNAs should be approximately 35-43 nucleotides, a value which is entirely consistent with the observed length distribution values in Fig. 6.

### 3.2 Minicircle-encoded gRNAs

Each kDNA network contains approximately 5000-10000 minicircles linked together by catenation. The size of the minicircle is species-specific and varies from 400 bp to 2500 bp. The minicircle molecule is organized, depending on the species, into one, two, or four conserved regions and a corresponding number of variable regions. The conserved region contains the origins of replication for both strands. In most kinetoplastid species there are multiple minicircle sequence classes, which are defined by the variable region sequences, within a single network. The number of different sequence classes in the African trypanosome, *T. brucei*, is over 300 (21).

The genetic role of this enigmatic molecule was finally uncovered when it was shown that the *L. tarentolae* D12 minicircle encodes a gRNA which potentially mediates the editing of sites 1-8 of the COIII mRNA (64). In *L. tarentolae*, each minicircle encodes a single gRNA located approximately 150 bp from the end of the conserved region and the 'bend' in the DNA (65) (Fig. 7). In *T. brucei*, each minicircle encodes at least three gRNAs which are located between 'cassettes' of 18 bp inverted repeats (66) (Fig. 7). However, four genes encoding redundant gRNAs mediating the editing of block 1 of CYb were shown to be localized outside the 18 bp inverted repeat cassettes in different *T. brucei* minicircles (67).

The extent of minicircle complexity in the UC strain of *L. tarentolae* was analysed by screening several large minicircle DNA libraries for different gRNAs by a process of negative colony hybridization selection (68). A total of 17 minicircle sequence classes was shown to exist, each of which was found to encode a gRNA (Tables 1 and 2). The editing role of all but two of these gRNAs was determined by comparison with known edited mRNA sequences. The relative abundance of the different minicircles was determined by dot blot analysis to vary from as few as 30 copies to as many as 3000 copies per network of 10000 minicircles. The relative abundance of the steady-state gRNAs encoded by these minicircles was measured.
Fig. 7 Genomic organization of kDNA minicircles from two trypanosomatid species, showing locations of the conserved region (black box), the DNA 'bend', and the gRNA genes (arrows show polarity). CSB-3 is a conserved 12-mer involved in replication initiation. The 18-mer inverted repeats flanking most of the gRNA genes in *T. brucei* are indicated by arrows.

by northern blot analysis and shown generally to have no correlation with the minicircle copy number. This indicates that steady-state gRNA abundance is determined by relative promoter strength or turnover rather than by a gene dosage effect.

In *T. brucei*, each 1 kb minicircle encodes three gRNAs usually situated within three cassettes of 18 bp inverted repeats (5'-'GAAATAAGATAATAGATA—~110 bp—'TATTTATTTTTATTTTT-3') in the same polarity (66) (Fig. 7). The gRNA transcripts can be capped *in vitro* by vaccinia virus guanylyl transferase, and therefore may represent primary transcripts, although this has not yet been experimentally demonstrated. Transcription appears to initiate at a purine within the sequence, 5'-'AYAYA-3', 32 bp from the upstream inverted repeat. As in the case of *L. tarentolae* gRNAs, the *T. brucei* gRNAs have non-encoded 3' oligo(U) tails of heterogeneous length. In *T. brucei*, the minicircle sequence heterogeneity is more than 10-fold larger than that in *L. tarentolae*, suggesting that the total gRNA complement comprises more than 700–900 different sequences. This large repertoire of gRNAs can only partially be explained by the existence of three pan-edited cryptogenes—*ND7*, *COIII*, and *MURF4*—which are only 5'- and internal-edited in *L. tarentolae* and *Crithidia*. To date, seven gRNAs have been reported for *ND7* in *T. brucei* (69), 10 for *ND8* (70), 11 for *A6* (MURF4) (70), five for *CR1* (G1) (37), and 17 for *COIII* (70), but in no case has a complete set of overlapping gRNAs yet been obtained. These gRNAs represent approximately 30% of the known edited sequence information in this species. Of the 50 identified gRNAs, 32 show some level of overlap with another gRNA, with the extent of overlap ranging from 2 to 52 nucleotides. In some cases, completely overlapping 'redundant' gRNAs were observed, differing in sequence but encoding the same editing information (70). This gRNA redundancy is probably the major factor contributing to the large number of gRNAs in this species. The reason for this is not known but may be
### Table 2 Guide RNA complexity in *L. tarentolae* UC and LEM125 strains

<table>
<thead>
<tr>
<th>Cryptogenes</th>
<th>UC + LEM125</th>
<th>UC + LEM125</th>
<th>Total (expected)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>maxicircle DNA</td>
<td>minicircle DNA</td>
<td></td>
</tr>
<tr>
<td>COII</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>COIII</td>
<td>0</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>ND7</td>
<td>2</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>Cyb</td>
<td>2</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>MURF2</td>
<td>2</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>MURF4 (A6)</td>
<td>0</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>RPS12 (G6)</td>
<td>1</td>
<td>7</td>
<td>8</td>
</tr>
<tr>
<td>ND8 (G1)</td>
<td>1*</td>
<td>9</td>
<td>9 (14)</td>
</tr>
<tr>
<td>ND9 (G2)</td>
<td>1</td>
<td>8</td>
<td>9 (17)</td>
</tr>
<tr>
<td>G3</td>
<td>2*</td>
<td>1</td>
<td>3 (6)</td>
</tr>
<tr>
<td>G4</td>
<td>1</td>
<td>9</td>
<td>10 (15)</td>
</tr>
<tr>
<td>ND3 (G5)</td>
<td>1*</td>
<td>5</td>
<td>6 (9)</td>
</tr>
<tr>
<td><strong>Total:</strong></td>
<td><strong>13</strong></td>
<td><strong>47</strong></td>
<td><strong>60 (83)</strong></td>
</tr>
</tbody>
</table>

* gND8-XII, a putative maxicircle-encoded gRNA with several mismatches.
* gG3-I, a putative maxicircle-encoded gRNA which was not detectable by northern and primer extension analysis.

related to the developmental regulation of editing that occurs during the complex life cycle of the African trypanosome. In this regard, the defective African trypanosomes, *T. equiperdum*, *T. evansi*, and *T. equinum*, which lack the insect stage of the life cycle, have homogeneous minicircle populations which encode three gRNAs (71, 72). However, in these species, the maxicircle DNA either has large deletions and rearrangements (73) or is absent, and editing does not occur.

The 3' terminus of gCyb-II from *L. tarentolae* (60) was mapped by S1 analysis. The majority of molecules had the 3' ends predicted from the alignment of gRNA with edited mRNA, but a low percentage of truncated molecules were also observed, which were suggested to be due to premature termination. Another method for analysis of gRNA 3' ends is to obtain sequences of gRNA–mRNA chimeric molecules (see below). By this method, evidence for a high proportion of 3' truncated gRNAs was obtained in the case of *T. brucei*, and the suggestion was made that truncation may occur during the editing process (69). Similar results were obtained with gRNA–mRNA chimeric molecules from *C. fasciculata*, and the suggestion was made that the apparent 3' to 5' progression of editing determined by the alignment of partially edited clones (47) may be an artifact of editing mediated by 3'-truncated gRNAs (74).

However, this is definitely not the situation in the case of the partially edited *L. tarentolae* Cyb transcripts analysed by Sturm and Simpson (75). The 3' ends of all
identified \textit{L. tarentolae} gRNAs from both the UC and the LEM125 strains were sequenced directly by RACE-PCR during the construction of a gRNA library (O. Thiemann and L. Simpson, unpublished; 40), and were found to have a slight amount of 3' end heterogeneity, but truncated molecules such as observed in \textit{C. fasciculata} were not observed.

### 3.3 Complete sets of overlapping gRNAs for the CYb, MURF2, COIII, MURF4, and RPS12 mRNAs in \textit{L. tarentolae}

Two overlapping maxicircle-encoded gRNAs mediate the 5'-editing of the CYb mRNA in \textit{L. tarentolae} (22), but the duplex anchor for gCYb-II would be only 5 base pairs. In \textit{C. fasciculata} (76), there are also two overlapping gRNAs for the editing of CYb. However, in this case, a stable duplex anchor for gCYb-II is created by the editing mediated by gCYb-I. If there had been a sequence error for the \textit{L. tarentolae} gCYb-I involving a single U residue in a string of 5 Us, then a stable 13-nucleotide \textit{L. tarentolae} gCYb-II anchor sequence would also be created by block I editing (G. Connell, personal communication).

Two maxicircle-encoded gRNAs are involved in the editing of the MURF2 mRNA in \textit{L. tarentolae}, gMURF2-I is unusual in that it only encodes a single U-addition.

Two minicircle-encoded gRNAs mediate the editing of the 5'-edited COIII RNA in \textit{L. tarentolae} (68). The anchor sequence for gCOIII-I is complementary to unedited mRNA sequence, whereas the anchor for gCOIII-II covers editing sites 5-8 of gCOIII-I.

Six overlapping minicircle-encoded gRNAs mediate the editing of the 5' pan-edited MURF4 mRNA in \textit{L. tarentolae} (68). The 3'-most gRNA (gMURF4-I) forms an anchor duplex with unedited sequence, but all the other gRNAs form anchors with edited sequences. There is a possibility that gMURF4-V forms a six base-pair anchor with unedited sequence, and therefore would create a second editing domain, but this must be verified by analysis of partially edited RNAs.

Eight overlapping gRNAs mediate editing of the RPS12 mRNA in \textit{L. tarentolae} (68) (Fig. 8). Seven of these are minicircle-encoded and one is maxicircle-encoded (gRPS12-VI). The initial gRNA in each of the three domains forms an anchor with unedited sequences, whereas all other gRNAs form anchors with edited sequences.

The precise overlapping of gRNAs in these three pan-edited cryptogenes in this species is strikingly economical and provides a simple explanation for the observed 3' to 5' polarity of editing within a domain (Figs 5 and 8).

Analysis of these gRNA-mRNA hybrids showed that the G-U base pairs are limited to regions of the edited mRNA upstream of the anchor sequence (68), and there are few in anchor duplexes. This suggests a role for the relatively weak G-U base pairs in allowing breathing of the edited mRNA-gRNA duplex block, which would permit formation of the adjacent upstream mRNA-gRNA anchor duplex. This would not, however, eliminate the possible need for an RNA helicase to de-
Fig. 8 Diagram of overlapping gRNAs for editing of RPS12 mRNA. The unedited mRNA sequences used for anchors by the initial gRNAs in each domain are indicated by black bars and the gRNA/mRNA anchor duplexes by straight lines. The genomic origins of the pre-edited mRNA and each gRNA are indicated by arrows.

stabilize the terminal gRNA-edited mRNA duplex to allow translation, and such an activity has recently been reported in mitochondrial extracts from T. brucei (77).

3.4 Sets of overlapping gRNAs for the G1–G5 mRNAs in the LEM125 strain of L. tarentolae

As discussed above, a complete set of productively edited mRNAs for G1–G5 was obtained from the LEM125 strain of L. tarentolae. A gRNA library was constructed for this strain (40), and gRNAs mediating the editing of these transcripts were selected by a process of negative selective hybridization, using DNA oligonucleotide probes for gRNAs from the UC strain and sequencing of remaining clones. By this method, 30 new minicircle-encoded gRNAs and five new maxicircle-encoded gRNAs were identified (Tables 1 and 2). Many of the gRNAs were found to overlap and the overlap was restricted to the anchor region, as was described previously for the MURF4 and RPS12 editing cascade in L. tarentolae UC strain. It is likely that a more extensive search would yield a complete set of overlapping gRNAs for these transcripts, and we therefore conclude that these transcripts are productively edited in this strain.

One example of a redundant gRNA was detected for the editing of block III of the ND3 transcript (40). This indicates that redundant gRNAs are not restricted to the African trypanosomes. However, it is clear that the abundance of redundant gRNAs is much lower in L. tarentolae than in T. brucei.

Two previously unassigned gRNAs from the UC strain of L. tarentolae—gLt19 and gB4 (68)—proved to be identical homologues of the gRNAs which mediate editing of G4 block III and ND3 block IX in the LEM125 strain (40). It is of some
interest that the minicircles encoding these two gRNAs represent the highest copy number minicircles in the UC strain, but are significantly lower copy number in the LEM125 strain.

3.5 Loss of minicircle-encoded gRNAs in an old laboratory strain of L. tarentolae

In the case of L. tarentolae, productive editing of transcripts of the G1–G5 cryptogenes was only observed in the recently isolated LEM125 strain (40). Attempts in our laboratory to RT-PCR amplify partially and fully edited RNAs from the UC strain, which has been in culture in various laboratories for over 50 years, were uniformly unsuccessful, except in the case of G5, which gave rise to a percentage of clones containing correctly edited block I sequence and incorrectly edited upstream sequences (see Section 6). Examination of the total gRNA complement of the UC strain (Table 2) led to the conclusion that this failure to properly edit these transcripts was most likely to be a result of the absence of specific minicircle sequence classes encoding gRNAs for these specific editing cascades. At least 32 new minicircle-encoded gRNAs for the editing of the G1–G5 transcripts have been shown to be present in the LEM125 strain and to be absent in the UC strain (Tables 1 and 2) (40). We have speculated that specific minicircle sequence classes encoding these gRNAs have been lost during the long cultural history of the UC strain, possibly due to mis-segregation of minicircles at division of the single kDNA network or to the selective amplification of minor sequence classes which occurs in the phenomenon of transkinetoplastidy (78). The loss of these genes is tolerated since the encoded components of complex I of the respiratory chain are apparently not required during this stage of the life cycle (79).

The loss of gRNA genes in the UC strain has created several 'pseudo-cryptogenes', which are transcribed, but the transcripts are not productively edited. The observed correlation of the loss of specific sets of gRNA genes with defects in editing of these transcripts provides the first in vivo evidence for the genetic role of gRNAs in RNA editing.

4. Models for the mechanism of RNA editing and the involvement of gRNAs

The sequence information for addition and deletion of Us is contained in the gRNA molecules, but the precise mechanism by which this information is transferred to the mRNA is still unresolved. The enzyme cascade model (22) (Fig. 9) was based on the existence of several enzyme activities in purified kinetoplast-mitochondria fractions of L. tarentolae—a 3' TUTase, an RNA ligase, and a specific RNA cleavage activity (54, 80). The initial event was proposed to be the hybridization of the 5' portion of the gRNA to the mRNA anchor sequence just downstream of the pre-edited region. Then a specific cleavage was postulated at the 3'-most mismatched
Fig. 9 Models for the mechanism of RNA editing in kinetoplastid mitochondria. PER, pre-edited region. Only U insertions are shown. The guiding a and g nucleotides in the gRNA are shown in lower case and circled. The gRNA shown is gCyb-1.

base on the mRNA, liberating an internal 3'-OH. The addition of a U residue to this 3'-OH was then hypothesized, which would form a base pair with a guiding A or G residue on the gRNA. The final event would be the religation of the 5' mRNA fragment with the 3' mRNA fragment. Another cycle of editing at the next upstream mismatched base would then ensue. Deletions were presumed to be due to exonuclease trimming of an exposed non-base-paired 3' U residue in the mRNA. Evidence for this model is mainly the existence of all three postulated enzymatic
activities in the mitochondrion, and the co-localization of several of these activities in ribonucleoprotein complexes in mitochondrial extracts from T. brucei (81).

The presence of a heterogeneous 3' oligo(U) tail on the gRNA presents a puzzle. We initially suggested that this tail could form a duplex structure with the pre-edited region which is mainly composed of G and A residues, and thereby increase the stability of the initial mRNA–gRNA hybrid (60). In the enzyme cascade model the role of the TUTase is both to add Us at the editing site and to add the Us to the 3' ends of the gRNAs. A more direct role for the oligo(U) tail was proposed in the transesterification model of editing (82, 83) (Fig. 9). In this model, the role of the TUTase is to maintain the oligo(U) tail of the gRNA for use as a donor of U residues to the editing site. The initial event is a cleavage–ligation attack by the 3'-OH of the gRNA at the mismatch, giving rise to a transient gRNA–mRNA chimeric molecule. This initial transesterification is activated by hybridization of the 3'-terminal U residues of the gRNA with the internal guide sequence of the gRNA in a manner similar to that occurring in the ribosomal RNA self-splicing intron of Tetrahymena (84–89; Chapter 1). A second transesterification between the 3'-OH of the upstream mRNA fragment and the next mismatch liberates the gRNA and produces the transfer of at least one U residue from the gRNA into the mRNA at an editing site (Fig. 9). Deletions of Us could occur either by the initial transesterification being 5' of the mismatched U residue with the second transesterification being 3' of that residue, or by an exonuclease trimming of the exposed 3' U residue after the initial transesterification at a normal site.

This mechanism is formally very similar to the reversal of RNA splicing of group II introns, which has been observed in in vitro models (84; Chapter 1). Evidence for this model mainly rests on the existence of the predicted gRNA–mRNA chimeric molecules, which will be discussed below.

5. Chimeric gRNA–mRNA molecules

A prediction of the transesterification model for editing is the existence of transient gRNA–mRNA chimeric molecules. Such molecules were isolated from L. tarentolae kinetoplast RNA by PCR amplification, using a 3' primer specific for the mRNA downstream of the pre-edited version and 5' primer specific for the gRNA (82). These chimeric molecules have gRNAs usually attached at normal editing sites by 8–26 U residues, with the downstream editing sites being completely edited. Chimeric molecules for ND7, COIII, and COII were observed. In the COII chimeras, the connecting nucleotides are Us and As, which is consistent with the fact that the COII gRNA is in cis at the 3' terminus of the mRNA.

The majority of gRNAs were found to be attached at the 3'-most editing sites or at sites with a large number of U additions. The 3'-most encoded nucleotides in the gRNA portion are in most cases consistent with the 3' ends predicted by the largest gRNA-edited mRNA duplex that can be formed. However, several examples of 3'-truncated gRNAs were observed in the case of ND7. Chimeric molecules for the pan-edited MURF4 cryptogene were also observed in T. brucei mitochondrial RNA
In this case, the majority of the encoded gRNA sequences are 3' truncated and the connecting oligo(U) sequence is more heterogeneous and shorter than in *L. tarentolae*.

It should be emphasized that the existence of these molecules is consistent with the transesterification model but does not prove it, since chimeric molecules could also be generated by adventitious ligation of the 3' end of the gRNA to the 3' mRNA fragment produced in the enzyme cascade model. The possibility of PCR-generated *in vitro* homologous recombination (90) is made unlikely by the fact that chimeric molecules from *T. brucei* have also been detected in non-amplified cDNA libraries (K. Stuart, personal communication) and as a minor high molecular weight band representing approximately 5% of the gRNA abundance in total mitochondrial RNA by northern blot analysis in *L. tarentolae* (82).

### 6. Generation of misedited patterns: random editing or misguiding?

It was initially shown for the pan-edited COIII cryptogene that editing proceeded generally 3' to 5' (91). Analysis of a large number of partially edited molecules for CYb and COIII in *L. tarentolae* indicated that editing also proceeds 3' to 5' within a single editing block (75). Almost all of the CYb partially edited RNAs showed a precise 3' to 5' polarity, but 42% of the COIII partially edited RNAs showed unexpected editing patterns at the junction regions. We have presented evidence that unexpected editing patterns are produced by misediting due to specific events of misguiding through the interaction of inappropriate gRNAs, or of appropriate gRNAs in an inappropriate fashion (47). The basic concept is that of a 'guiding frame': misediting represents 3' to 5' editing occurring in an incorrect guiding frame. We have proposed four possible misguiding mechanisms and have presented several examples of each type (Fig. 10). Chimeric molecules consisting of misedited mRNAs attached to heterologous gRNAs which had the potential of guiding the editing events were also observed. Misguiding would represent mechanistically correct editing, but misedited mRNAs must always be corrected by in-frame editing mediated by the appropriate gRNA to obtain the mature edited mRNA. It is of course also possible that this mechanism may not account for all observed misedited sequences and that a certain amount of stochastic editing does occur in addition.

The preferential localization of misediting to junction regions can be explained as a natural consequence of the misguiding hypothesis. The scattered clusters of misedited sequences observed at a low frequency in the *T. brucei* CR1 partially edited RNAs could be due to the greater diversity of redundant gRNAs in this species, which may make pan-editing inherently inaccurate (37).

Several illustrative examples of misediting were recently obtained for the ND3 (G5) cryptogene in the UC strain of *L. tarentolae* (40, 92). Analysis of clones of partially edited G5 transcripts from the UC strain showed a subset that possessed a
correctly edited block I sequence followed by a variety of misedited upstream sequences (Fig. 11). The presence of a maxicircle-encoded gRNA in the UC strain explains the correct editing of block I. The gRNA for block II in the LEM125 strain is minicircle-encoded and is absent in the UC strain. Several examples of misedited sequences extending from a correctly edited block I sequence were observed, in which the misediting appeared to be mediated by non-cognate gRNAs (92). In a few cases, the misediting created false anchor sequences for additional non-cognate gRNAs which further extended the misediting. Another group of clones with polyadenylation sites up to six nucleotides from the anchor for block I editing showed misediting of block I also, suggesting the possibility of some minimal space requirements for the assembly of an editing complex on the 3' end of a pre-edited mRNA.

An analysis of partially edited RNAs from the CYb and COIII genes of T. brucei showed a high abundance of misedited patterns at junction regions (93). These
Fig. 11 Misediting of G5 transcripts in UC strain of *L. tarentolae*. (a) Correct editing of block I in both strains by a maxicircle-encoded gRNA (gND3-I), and incorrect editing of the upstream sequences in the UC strain due to an absence of the cognate minicircle-encoded gRNAs. (b) Example of one partially edited G5 RNA from the UC strain which has correct editing of block I and misediting of upstream sequence due to the sequential actions of two non-cognate misguiding gRNAs. The correctly edited block I is boxed, and the gRNAs are shown above and below the edited RNA sequence. The mature edited LEM125 G5 sequence and the cognate gRNAs are shown for comparison. (Reprinted from 40 with permission.)
results were interpreted as evidence that editing does not proceed strictly 3' to 5', but is stochastic in nature—in other words, that editing is essentially completely random within a domain, and that formation of the correct duplex with the gRNA 'freezes' editing in that region (93). In this model, the secondary structure of the pre-edited mRNA itself determines the sites of editing by presenting a single-stranded loop region to a single strand-specific endonuclease (80, 94). Subsequent events could involve either the enzyme cascade steps or the transesterification steps for the U additions and deletions. gRNA-mRNA chimeric molecules could also be generated in this model by ligation of the 3' end of the gRNA to the mRNA at the cleavage site, but these would represent aberrant products unless a second cleavage-ligation occurred, releasing the shortened gRNA (Fig. 9).

The 'dynamic interaction' model (95) is also an attempt to explain the existence of misedited sequences as a normal consequence of editing. The model suggests a series of progressively more stable, but incompletely base-paired, mRNA-gRNA interactions in the junction region. This model is essentially identical to one type of misguiding in which there is a loopout of either the mRNA or the gRNA leading to a loss of guiding frame (47). Editing of sites would occur multiple times and overall editing would not proceed strictly 3' to 5'. The model proposes a progressive realignment of gRNA-edited mRNA sequences to form the final mature edited mRNA, and is compatible with both the enzymatic and the transesterification mechanisms of editing.

The misguiding hypothesis of Sturm et al. (47) can explain the origin of a substantial number of misedited sequences in L. tarentolae, and a knowledge of the total gRNA complement in T. brucei should allow a test of the validity of this and other hypotheses for T. brucei.

7. Enzymatic activities possibly associated with RNA editing

TUTase and RNA ligase activities were first reported in total cell extracts of T. brucei (74). Mitochondrial TUTase and RNA ligase activities were demonstrated in L. tarentolae (54). The TUTase adds U residues to the 3' OH of most substrate RNAs, but the mitochondrial gRNAs represent exceptionally good substrates. The activity is inhibited by heparin, and is specific for UTP. A TUTase activity was also observed in T. brucei kinetoplast-mitochondrial fractions isolated by Percoll density gradient centrifugation (81, 96). The mitochondrial RNA ligase activity in L. tarentolae is ATP-dependent. Both the TUTase and the RNA ligase activities were solubilized, in the case of L. tarentolae, with Triton X-100.

A ribonuclease activity in a 100000 g supernatant of a Triton lysate of a mitochondrial fraction from L. tarentolae was shown to be activated by incubation with heparin or by predigestion of the lysate with proteinase K (80). In vitro-transcribed CYb mRNA is cleaved at several sites within putative single-stranded regions of the mRNA. The major cleavage occurs two nucleotides upstream from the initial
ENZYMATIC ACTIVITIES POSSIBLY ASSOCIATED WITH RNA EDITING

Fig. 12 Localization of sites of cleavage within the synthetic CYb mRNA molecule. pNB2 RNA is a fragment of pre-edited CYb mRNA with some vector sequence at the 5' end. At the left side is a sequencing ladder of intact pNB2 RNA, obtained using a [32P]-labeled oligonucleotide primer. Sites of editing and number of Us added in the mature RNA sequence are shown by arrows. The pNB2 RNA was incubated in the presence of heparin with mitochondrial extract from L. tarentolae, which was pre-digested with proteinase K, and the two major cleavage fragments were gel-isolated. A sequencing ladder of the fragments is shown on the right, with the nucleotides adjacent. The sites with nucleotides in all lanes are interpreted as sites of cleavage and are indicated by arrows on the left and by N. on the right. (Reprinted with permission from 80.)

The cleavage activity is inhibited by SDS or extraction with phenol/chloroform and therefore probably represents a protease-resistant protein, rather than an RNA ribozyme. The activity was sized between 10 kDa and 30 kDa by ultrafiltration. The results of micrococcal nuclease digestion were equivocal, in that digestion inhibits the protease-induced activity but not the heparin-induced activity. A similar activity was observed in T. brucei mitochondrial extracts, using synthetic CYb, COII, and COII mRNAs (94). Pre-edited CYb RNA is specifically cleaved at the identical site as in L. tarentolae, but mature edited RNA is not affected. A specific cleavage just downstream of the PER in the COII mRNA and multiple cleavages within the large PER of the pan-edited COIII mRNA were also observed. Since it was also observed that the single-strand-specific nuclease, mung bean nuclease, cleaved the CYb RNA at identical locations, the suggestion was
made that the mitochondrial activity represents a single strand-specific endonuclease. However, in all cases, the cleavage activity had to be activated by inhibition of the TUTase with heparin, digestion with proteinase K, or depletion of UTP by preincubation of the isolated mitochondria.

The activation effect is not a definitive proof of the enzymatic mechanism since it could possibly also be explained in terms of the transesterification model for editing. In group I splicing, hydrolysis at normal sites of transesterification occurs in the absence of the guanosine residue which provides the attacking 3'-OH group (Chapter 1). The observed cleavages of pre-edited mRNAs could represent hydrolysis catalysed by a protease-resistant catalytic core in the absence of activated gRNA (80).

8. **In vitro formation of gRNA–mRNA chimeric molecules**

Several reports have shown the presence of an activity in mitochondrial extracts that promotes the formation of gRNA–mRNA chimeric molecules *in vitro*. Harris and Hajduk (97) used labeled synthetic gRNA or mRNA from *T. brucei* to monitor the formation of chimeric molecules for the CYb cryptogene by gel electrophoresis. Synthetic pre-edited CYb mRNA was used, together with a synthetic gRNA covering the first three editing sites. The sequence of the gRNA was taken from the *L. tarentolae* gCYb-I gRNA. Harris and Hajduk (97) were unable to PCR-amplify the *in vitro* products and therefore attempted to directly sequence 5'-end-labeled chimeric molecules. Only very limited direct sequence information was obtained from the *in vitro* products and the site of attachment of the gRNA remained ambiguous.

Koslowsky *et al.* (98) employed PCR amplification to analyse chimeric molecules formed with synthetic A6 (MURF4) pre-edited mRNA and synthetic gA6–14 gRNA from *T. brucei*. However, the amplified *in vitro* chimeras were unusual in that they lacked a stretch of U residues linking the gRNA to the mRNA and most of the gRNAs were truncated at variable sites at the 3' end, suggesting that the process was aberrant in some respect. In both studies, specificity was demonstrated by showing that a variety of heterologous RNAs did not form chimeras with added gRNA. In addition, a requirement for a gRNA 3'-terminal hydroxyl group was demonstrated and addition of protease-sensitive mitochondrial extract was required for chimeric molecule formation.

Blum and Simpson (99) have obtained similar results in the case of *L. tarentolae*. Synthetic pre-edited messenger RNA and synthetic gRNA for the ND7 cryptogene from *L. tarentolae* form chimeric molecules upon incubation in the presence of an extract from sonicated mitochondria. These chimeric molecules consist of the gRNAs covalently linked to the mRNAs by short oligo(U) tails at normal editing sites in most cases (Fig. 13). Unlike the *in vivo* chimeras in steady-state kinetoplast RNA, the *in vitro* chimeras showed no editing of downstream editing sites. The
formation of chimeras requires ATP and is dependent on the formation of a gRNA–mRNA anchor duplex 3' of the pre-edited region, as shown by in vitro mutagenesis of the mRNA and restoration of activity by compensatory base changes in the gRNA. mRNA sequences 3' and 5' of the pre-edited region also affect the efficiency of the chimera-forming activity.

One possible explanation for the synthesis of aberrant chimeric molecules in vitro is a failure to rejoin the two separated mRNA fragments after chimera formation. Alternatively, correctly edited products may actually be formed during the incubation, but may undergo rapid hydrolysis (80). This would be consistent with the observed lack of chimeras with gRNAs attached at editing site 1, which is the preferred attachment site for in vivo chimeras (82). The remaining stable chimeras would then represent aberrant products accumulating during the incubation. Nevertheless, the observed in vitro formation of chimeras may still accurately reflect the initial step of RNA editing, consistent with the transesterification model for RNA editing. Alternatively, chimeric molecules could be created by site-specific cleavage (81, 94) and adventitious ligation of the 3' oligo(U) tail of the gRNA to the cleaved mRNA 3' fragment, in line with the enzyme cascade model. The development of an accurate and complete in vitro editing system will be required to definitively distinguish between these models, but the availability of an in vitro system for the formation of chimeric gRNA–mRNA molecules should already allow a precise dissection of the sequence requirements for chimera formation and also a fractionation of the extract components required for this activity.

9. Ribonucleoprotein complexes containing putative components of the editing machinery

A variety of mitochondrial ribonucleoprotein (RNP) complexes containing putative components of the editing machinery have been reported. Two classes of gRNA-containing RNP complexes in a mitochondrial extract from T. brucei can be detected.
by sedimentation in glycerol gradients: a 19S complex which contains gRNA, TUTase, RNA ligase, and chimera-forming activity, and a 35S complex which has in addition pre-edited RNA, but lacks tightly bound TUTase (81). Pollard et al. (81) suggested that the cosedimentation of the RNA ligase and chimeric-forming activity indicates that the mechanism for chimeric formation does not involve trans­esterification but rather cleavage–ligation. Gel retardation analysis has also been used to detect several mitochondrial RNP complexes from T. brucei which interact with exogenous synthetic gRNAs, but the function of these complexes remains unknown (100–102).

Peris et al. (103) have described two classes of RNP complexes in a mitochondrial extract from L. tarentolae. The 'T-complexes' contain gRNAs and mRNAs and are operationally defined as being labeled with [α-32P]UTP by an endogenous TUTase activity. These complexes sediment at 10-13S in glycerol and migrate in native gels as about six bands. T-complexes may represent gRNA-maturation complexes, in regard to maintenance of the 3' oligo(U) tails of the gRNAs. The 'G-complexes' also contain gRNAs, but sediment at 25S in glycerol, and exhibit an in vitro RNA editing-like activity, which is not found in the T-complex region of the gradient. This RNA editing-like activity involves the incorporation of uridine residues into the pre-edited region of a synthetic mRNA (104) (see Section 10).

10. In vitro RNA editing-like activities

Seiwert and Stuart (105) have described an in vitro activity in a mitochondrial extract from T. brucei which accurately removes uridines from the first editing site of the A6 (MURF4) pre-edited mRNA. This deletion activity is dependent on the presence of ATP and exogenous guide RNA for this editing block (gA6[14]). Mutations in both the gRNA and the mRNA editing site indicate strongly that the number of Us deleted in vitro is controlled by base-pairing interactions. This represents, the first in vitro evidence confirming the predicted genetic role of gRNA in specifying the sequence information for editing, at least in the case of U-deletions. However U-additions were not observed in this in vitro system, raising the possibility that they have a different mechanism than U-deletions.

Frech et al. (104) showed that a mitochondrial extract from L. tarentolae could direct the incorporation of U residues derived from [α-32P]UTP within synthetic CYb mRNA. The U-incorporation is imprecise but is confined to the pre-edited region. No direct evidence for the involvement of endogenous gRNAs was obtained, and, interestingly, the addition of exogenous gRNAs (or other nonspecific RNAs) inhibits the incorporation. However, specific inhibition by digestion of the extract with micrococcal nuclease suggests a requirement for some type of endogenous RNA. A low level of incorporation of C residues was found to occur at the same sites as U residues. This activity sediments in glycerol at 20–25S and is correlated with the presence of the G complexes described above.

Further analysis of these in vitro systems exhibiting editing-like activities should allow a distinction between the enzymatic and transesterification mechanisms of
evolution of RNA editing, and may prove useful in an eventual dissection and reconstitution of editing activities.

11. Evolution of RNA editing in kinetoplastid protozoa

Phylogenetic trees of the kinetoplastid protozoa have been constructed in several laboratories using nuclear small rRNA sequences (106–108). The trees were rooted using the Euglena rRNA sequence as an outgroup. As predicted by classical taxonomy, the bodonids/cryptobiids form a sister group to the trypanosomatids. In the trypanosomatids, the earliest diverging species are the African and South American Trypanosoma, and the most recently diverging species comprise a monophyletic clade consisting of Leishmania, Crithidia, Endotrypanum, and Leptomonas. In the central portion of the tree are several monogenetic (one host) genera such as Blastocrithidia and Herpetomonas and the digenetic (two hosts) Phytomonas. The extent of editing of three cryptogenes—MURF4, ND7, and COIII—was analysed for representative trypanosomatid species (106, 108) (Fig. 14). The most parsimonious interpretation of the results is that pan-editing is a primitive trait and that several times during the evolution of these cells, pan-edited genes were substituted by moderate- or 5’-edited genes (109). In one case, a fully edited COIII gene was apparently substituted for a pan-edited ancestral gene (108).

Two strains of a cryptobiid, T. borreli, were also analysed for editing of mitochondrial genes (19, 20). The kDNA in this cell was found to consist of two molecular species, as in the case of the trypanosomatids. A 40–80 kb circular component contains structural genes organized in a different order than in the maxicircle DNA of trypanosomatids. Several of these were shown to be cryptogenes, transcripts of which are edited, and in some cases pan-edited. A pan-edited RPS12 gene and a novel 5’- and 3’-edited CYb gene were identified, and the edited RNAs sequenced. COII and COIII are unedited in T. borreli. Putative gRNAs were shown to be encoded as tandemly organized 1 kb units in 270 kb circular molecules (19). These results indicate that gRNA-mediated pan-editing can be traced back to an ancestor of the entire kinetoplastid lineage.

Another apparently primitive trait, at least in the trypanosomatid lineage, is the presence of an extremely large gRNA repertoire with extensive redundancy. In the more recently evolved clade represented by Leishmania and Crithidia, the gRNA repertoire is much more limited, as discussed previously.

We have speculated, on the basis of the observed culture-induced loss of gRNA genes in the UC strain of L. tarentolae and of the distribution of editing patterns in various species, that the evolution of RNA editing in the kinetoplastids involves a replacement of the original pan-edited cryptogenes by 5’-edited or unedited homologues by cDNA copies of partially edited RNAs (108, 109) (Fig. 15). The selective pressure for this retroposition gene replacement could have been the loss of entire gRNA gene families by the loss of the corresponding minicircles. Those cells which had undergone a gene replacement of a pan-edited gene with a partially edited gene would survive the loss of gRNAs. It is possible that all original maxicircle
Most parsimonious phylogenetic tree
(Small subunit rRNA)

Genes

Fig. 14 Phylogeny of kinetoplastid RNA editing. Aligned sequences of 18S rRNAs were used to reconstruct a consensus phylogenetic tree using maximum parsimony. Bootstrap values are indicated for each node. The tree was rooted using the *Euglena gracilis* sequence as an outgroup. A diagram of the extent of editing of eight genes is shown on the right. Dark boxes indicate pan-edited regions, white boxes indicate non-edited regions, and grey boxes indicate a lack of information. The lack of a box indicates an absence of any information on the editing of that gene. All genes are shown 5' to 3'. 5' edited sequences are shown at the 5' end of the RNAs, although there is usually some unedited upstream sequence.
Genes were pan-edited GA-rich cryptogene skeletons, and that mature edited RNAs have replaced several of these, giving rise to the presently unedited genes such as ND1, ND4, ND5, and COI.

It is clear that complex gRNA-mediated RNA editing was present in an ancestor of the kinetoplastid flagellate lineage, but the exact time and the mode of origin of this phenomenon remain unclear. Analysis of additional lower eukaryotic cells, and the eubacterial cells that are the likely ancestors of the protomitochondrion endosymbiont, for the presence of the U-insertion/deletion editing should provide information on these issues.

Also, determination of the precise mechanism of this type of editing should provide some constraints for evolutionary speculation. For example, if the mechanism of RNA editing in kinetoplastid mitochondria turns out to involve transesterification, this would suggest that ribozyme-mediated RNA self-splicing or even mRNA splicing shared a common evolutionary history. In this view, introns and
gRNAs may both represent relics of the same primitive genetic systems which dealt with the information management of RNA molecules. Modern RNA editing in kinetoplastids would represent a remnant of a primitive mechanism for creating functional mRNA sequences which has been maintained in the mitochondria of these cells. On the other hand, if the mechanism turns out to be cleavage-ligation, then an alternative hypothesis would become viable, in which editing evolved in the mitochondrion of these cells and represents an ancient but derived trait (110-112).

12. Note added on proof

A brief description of significant progress in the field of insertion-deletion RNA editing during the production of this book is provided here.

(i) Putative intermediates in a gRNA-mediated U-deletion/insertion in vitro system, using the A6 pre-edited mRNA and a mitochondrial extract from T. brucei, were detected by gel analysis (113–115). The intermediates represented the A6 mRNA cleavage and ligation products predicted by the enzyme cascade model for editing at sites 1 and 2; gRNA-mRNA chimeric molecules were also detected, but evidence was presented that these molecules represented byproducts of the editing reaction.

(ii) A stereochemical investigation of an in vitro U-insertion activity in L. tarentolae mitochondrial extracts also provided results compatible with an enzyme cascade model and incompatible with models involving the transfer of Us from the 3' end of gRNAs to an editing site (116). Use of a primer extension assay with a pre-edited mRNA mutated in the anchor sequence showed that this activity was independent of both endogenous or exogenous gRNA; this activity was, however, dependent on the secondary structure of the pre-edited mRNA (117). A similar primer extension assay was used to show a U-insertion reaction in site 1 of the ND7 mRNA, which was mediated by the number of guiding nucleotides in the cognate gRNA (118). Blockage of the 3' end of the gRNA had little effect on the U-insertion activity, a result also compatible with an enzyme cascade model in which the inserted Us are derived from UTP. These results led to a model in which Us are first added to the 3' end of the 5' cleavage fragment of the pre-edited mRNA by a TUTase-like activity, and the oligo(U) 3' overhang not base paired to guiding nucleotides in the gRNA is then removed by a nuclease and the ends are then religated. An imprecise removal of the unpaired Us will lead to the observed gRNA-dependent unguided or misedited U-insertions at this site.

(iii) An analysis of mitochondrial lysates from L. tarentolae and T. brucei led to the identification of a 20S complex containing RNA ligase activity but lacking gRNA. gRNA was found to be localized within a 10S complex and also within a series of heterodisperse complexes that sedimented from 10S to over 30S. The suggestion was made that these two classes of RNP complexes interact in the editing reaction (119).

(iv) Construction of gRNA libraries from two strains of L. tarentolae allowed an analysis of the 3' ends (120). This analysis showed a remarkable homogeneity of the
3'-uridylylation sites, which could be due either to transcription termination or 3'-end processing.

(v) Minicircle-encoded gRNAs were isolated from *Crithidia fasciculata* (121) and from two strains of *Trypanosoma cruzi* (122), and the genomic organization of the gRNA genes analyzed. In *C. fasciculata*, each 2.5 kb minicircle contains a single gRNA gene in one of the two variable regions situated at a constant distance from the DNA bend. Five gRNAs were identified and all belonged to minor minicircles representing less than 2% of the total minicircle DNA. More than 90% of the kinetoplast DNA was composed of a single minicircle sequence class encoding an unidentified gRNA. These results suggest that the copy number of specific minicircle sequence classes can vary dramatically without an overall effect on the RNA editing system. In *T. cruzi*, each 1.45 kb minicircle encodes four gRNAs situated within the four variable regions. Multiple examples of redundant gRNAs were identified, which encode the same editing information but have different sequences. Extensive sequence polymorphisms, mainly transitions, within the variable regions of homologous minicircles from different strains were shown to account for the known lack of cross hybridization of the kinetoplast DNA from the strains.

(vi) Five gRNAs were identified from *T. borreli* (123) and were shown to be encoded in a 180 kb circular molecule, rather than in minicircles as in the trypanosomatids; structural genes and cryptogenes are encoded in an 80 kb circular molecule, which is the homologue of the maxicircle. The gRNAs are shorter than those from trypanosomatids and possess nonencoded oligo(U) sequences at the 3' and also at the 5' ends. The origin and role of the 5' oligo(U) sequence is unknown.

References


de la Cruz, V., Simpson, A., Lake, J., and Simpson, L. (1985) Primary sequence and


REFERENCES


