Available online at www.sciencedirect.com
SCIENCE DIRECT®
Molecular & Biochemical Parasitology 134 (2004) 183–191
MOLECULAR & BIOCHEMICAL PARASITOLOGY

Gene synteny and evolution of genome architecture in trypanosomatids

Elodie Ghedin[a,b], Frederic Bringaud[c], Jeremy Peterson[a], Peter Myler[d], Matthew Berriman[e], Alasdair Ivens[e], Björn Andersson[f], Esteban Bontempi[f,g], Jonathan Eisen[a], Sam Angiuoli[a], David Wanless[a], Anna Von Arx[a], Lee Murphy[e], Nicola Lennard[e], Steven Salzberg[a], Mark D. Adams[a,1], Owen White, Neil Hall[e], Kenneth Stuart[d], Claire M. Fraser[a], Najib M.A. El-Sayed[a,b,*]

[a] *Parasite Genomics, The Institute for Genomic Research, 9712 Medical Center Dr., Rockville, MD 20850, USA*
[b] *Department of Microbiology and Tropical Medicine, George Washington University, Washington, DC 20052, USA*
[c] *Laboratoire de Parasitologie Moléculaire, Université Victor Segalen de Bordeaux II, Bordeaux 33076, France*
[d] *Seattle Biomedical Research Institute, Seattle, WA 98109-1653, USA*
[e] *The Wellcome Trust Sanger Institute, The Wellcome Trust Genome Campus, Hinxton CB10 1SA, England, UK*
[f] *Center for Genomics and Bioinformatics, Karolinska Institute, Stockholm, Sweden*
[g] *Instituto Nacional de Parasitologia Dr. M. Fatala Chaben, ANLIS, Buenos Aires 1063, Argentina*

Received 30 September 2003; accepted 12 November 2003

Abstract

The trypanosomatid protozoa *Trypanosoma brucei*, *Trypanosoma cruzi* and *Leishmania major* are related human pathogens that cause markedly distinct diseases. Using information from genome sequencing projects currently underway, we have compared the sequences of large chromosomal fragments from each species. Despite high levels of divergence at the sequence level, these three species exhibit a striking conservation of gene order, suggesting that selection has maintained gene order among the trypanosomatids over hundreds of millions of years of evolution. The few sites of genome rearrangement between these species are marked by the presence of retrotransposon-like elements, suggesting that retrotransposons may have played an important role in shaping trypanosomatid genome organization. A degenerate retroelement was identified in *L. major* by examining the regions near breakage points of the synteny. This is the first such element found in *L. major* suggesting that retroelements were found in the common ancestor of all three species.

Keywords: Synteny; Genome architecture; Evolution; *Trypanosoma*; *Leishmania*; Retroelement

1. Introduction

Kinetoplastid protozoan parasites belong to a distinct evolutionary lineage of eukaryotes [1]. Within the kinetoplastids, members of the trypanosomatid family include etiological agents of several tropical diseases such as African sleeping sickness (*Trypanosoma brucei*), Chagas' disease (*Trypanosoma cruzi*), and cutaneous leishmaniasis (*Leishmania major*). According to the most recent studies, *Trypanosoma* and *Leishmania* last shared a common ancestor between 400 and 600 Ma [2–4], their divergence predating by far the emergence of mammals some 165 Ma [5]. Together, these three pathogens are responsible for more than 20 million human infections annually. They employ different immune evasion strategies with *T. brucei* undergoing antigenic variation of the surface coat [6], whereas *L. major* and *T. cruzi* adopt an intracellular lifestyle and invade host cells. *Leishmania* species exclusively elect residence within the parasitophorous vacuole of macrophages, whose role is to ingest and kill invaders. The genetic basis for these differences in parasitic modes is unknown.

In a first large-scale comparative genomics study in trypanosomatids, we analyzed the level of synteny (conservation of gene order) among these three organisms by

Abbreviations: BAC, bacterial artificial chromosome; GSS, genome survey sequence; CDS, coding sequence; Ma, million years ago; LTR, long terminal repeat; DIRE, degenerated *Ingi*/L1Tc-related element; SIRE, short interspersed repetitive element; CACK, calcium-activated potassium channel; DCP, dipeptidyl carboxypeptidase; PDK1, 3-phosphoinositide-dependent protein kinase-1; LRRP, leucine rich repeat protein; PTGT-4, putative glucose transporter

* Corresponding author. Tel.: +1-301-795-7525; fax: +1-301-838-0208.
E-mail address: nelsayed@tigr.org (N.M.A. El-Sayed).
[1] Present address: Case Western Reserve University, OH 44106, USA.

0166-6851/$ – see front matter © 2003 Elsevier B.V. All rights reserved.
doi:10.1016/j.molbiopara.2003.11.012

looking at homologous chromosomal segments. Analysis of synteny allows the close examination of the selective and mutational forces that act on chromosomal and genome structure. In addition, comparative gene annotation provides substantial aid in the assignment of orthology of genes across species since genes found in similar locations are frequently orthologous. We report in this paper the existence of a strong conservation of gene order in trypanosomatids and evidence for selective forces that appear to maintain the large directional clusters observed in these organisms [7–9]. Where differences between the species exist, interruption of synteny and genome rearrangement events are punctuated by the presence of retrotransposon-like elements or members of a gene family that has been shown to contain hot spots for retrotransposon insertion. The data suggest that retrotransposons likely played an important role in shaping trypanosomatid genome organization and evolution.

2. Materials and methods

2.1. Strains sequenced

The trypanosomatid reference strains being sequenced as part of multi-center collaborations are *T. brucei*, stock TREU927/4 single VAT derivative GuTat10.1; *L. major* MHOM/1L/81/Friedlin; and *T. cruzi*, CL Brener.

2.2. Genome assemblies and chromosomes

2.2.1. Leishmania major

The sequences used for this study comprise a 119 kb contig from *L. major* chromosome 12 (*Lm*Chr12; GenBank accession AL390114), a 200 kb contig from chromosome 15 (*Lm*Chr15; accession AL160371), and a 70 kb contig from chromosome 26 (*Lm*Chr26; accession AL160493). All three chromosomes are presently being sequenced using the whole chromosome shotgun approach. *L. major* chromosome 1 (*Lm*Chr1) is a 285 kb chromosome [7,10].

2.2.2. Trypanosoma brucei

The *T. brucei* genomic sequences represented in the alignments are portions of the 3 Mb chromosome IX (*Tb*ChrIX) and of the 1.05 Mb chromosome I (*Tb*ChrI) [9]. The *Tb*ChrIX portion is a contig from three bacterial artificial chromosome (BAC) inserts, RPCI93-3C4 (GenBank accession AC013485), RPCI93-1L12 (accession AC013484) and RPCI93-28G16 (accession AC092736). The first two BACs (RPCI93-3C4 and RPCI93-1L12) were selected by searching a database of end sequences from the *T. brucei* BAC library RPCI93 using all annotated genes of *Lm*Chr1, followed by DNA fingerprinting. The third BAC (RPCI93-28G16) was selected based on an end-sequence overlap with BAC RPCI93-1L12. The sequence assembly from the three BACs is 375,625 bp.

2.2.3. Trypanosoma cruzi

The *T. cruzi* genome segment homologous to *Lm*Chr1 was obtained by sequencing two small overlapping BACs (12A5, 14 kb in size, GenBank accession AC090077; and Tc3-4C6, 22.5 kb, accession AC094018). BAC 12A5 was initially selected by screening a *T. cruzi* BAC library with a nucleotide probe amplified from a genome survey sequence (GSS) clone with similarity to *Lm*Chr1. Tc3-4C6 was selected from the Tc3 *T. cruzi* BAC library based on an end-sequence overlap with BAC 12A5. These *T. cruzi* BACs were assigned to chromosome 8 (as defined in [11]) by hybridizing a *T. cruzi*-specific probe to genomic DNA separated by pulsed-field gel electrophoresis. Two large *T. cruzi* scaffolds generated in the assembly of the whole genome shotgun data were found to be homologous to the *Tb*ChrIX portion. Tc_scaffold_A is a scaffold of 144,204 bp in size comprised of 21 ordered and oriented contigs. Because it overlaps with the *Tc*Chr8 portion described above, only 119,204 bp are represented in the alignment. Tc_scaffold_B is 189,376 bp in length and comprised of 14 oriented and ordered contigs. A 174,376 bp region is represented in this alignment. The 135 kb portion of *T. brucei* chromosome I was aligned with a 105,864 bp region of *T. cruzi* chromosome 3 (AC137988.2).

2.3. Gene finding

Trained versions of GLIMMER [12] were the main gene finders used to predict coding regions in all assemblies. In order to determine homologies between assemblies, predicted coding sequences (CDSs) from one assembly were searched by BLASTX or BLASTP [13] against the nucleotide sequence or putative proteins of the corresponding assembly.

3. Results and discussion

To examine the level of synteny among the trypanosomatids, we identified *T. brucei* and *T. cruzi* chromosome segments that are homologous to the first fully sequenced chromosome from this order, *L. major* chromosome 1 (*Lm*Chr1). *Lm*Chr1 genes are arranged into two directional gene clusters, transcribed in opposite directions, separated by a 1.6 kb strand-switch region [7]. These alignments reveal that the level of synteny among the three organisms is extensive (Fig. 1a). Of the 79 CDSs reported on *Lm*Chr1 [7], we found homologs (possibly orthologs) for >80% of these on a 272.5 kb portion of *T. brucei* chromosome IX (*Tb*ChrIX) (Fig. 1a). Three CDSs not originally reported in *Lm*Chr1 were subsequently labelled as “conserved hypothetical” due to their homology to hypothetical proteins in *T. brucei*, bringing the total number of homologous genes between *Lm*Chr1 and *Tb*ChrIX to 65. In addition to the large segment inversion apparent between *Lm*Chr1 and *Tb*ChrIX, there are only two gene inversions and four gene duplications (or loss of a copy) in that section of the alignment. The *Tb*ChrIX fragment used in our analysis encompasses genes spanning

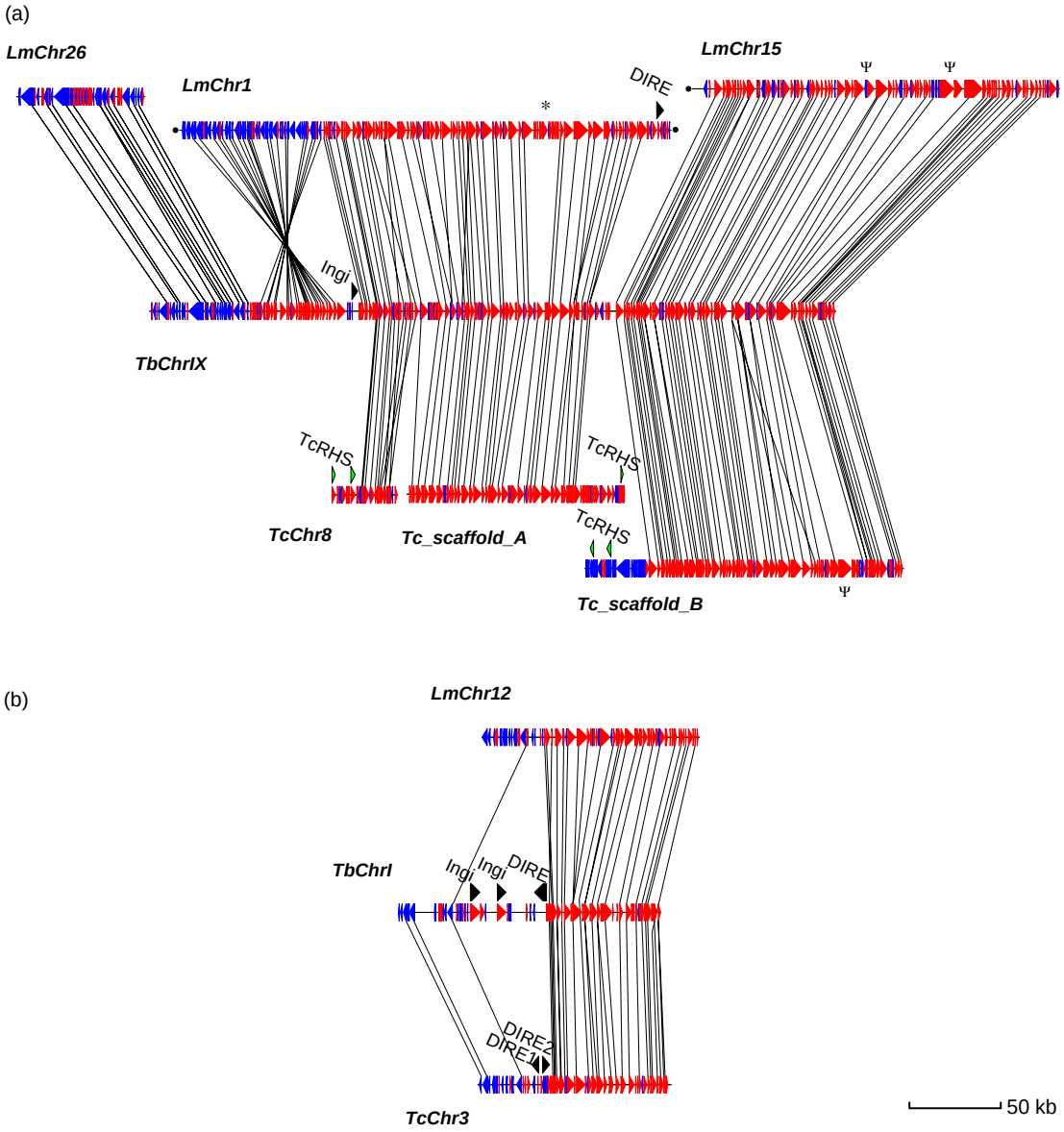

Fig. 1. Alignments of *T. brucei*, *T. cruzi* and *L. major* homologous genomic regions. Directionality of the gene clusters is represented by the color of the CDSs, red for one strand and blue for the other. Black arrows represent retrotransposon-like elements (*ingi* and DIRE). The green arrows indicate the location of *RHS* genes. Homologous genes are connected by a thin black line. (a) A 375 kb assembly of *T. brucei* chromosome IX (*Tb*ChrIX) was aligned to the 285 kb *L. major* chromosome 1 (*Lm*Chr1), a 70.3 kb portion of *Lm*Chr26 and a 200 kb portion of *Lm*Chr15. Two *T. cruzi* BACs, assigned to chromosome 8, make up a 36 kb portion of homology (*Tc*Chr8). Another two *T. cruzi* scaffolds make up the rest of the homologous region. The full black circles represent telomeric ends. The symbol asterisk (∗) represents a region of interruption of synteny. The symbol psi (Ψ) represents the presence of unique large genes (i.e. not part of the synteny). (b) A 135 kb portion of *T. brucei* chromosome I (*Tb*ChrI) was aligned with a 119 kb assembly of *L. major* chromosome 12 (*Lm*Chr12) and a 105 kb region of *T. cruzi* chromosome 3 (*Tc*Chr3).

all of *Lm*Chr1 as well as the beginning of two other *L. major* chromosomes. The extension beyond the corresponding extremities of *Lm*Chr1 reveals the presence of homologs to *Lm*Chr26 CDSs on one end, and CDSs from *Lm*Chr15 on the other (Fig. 1a). It is worth noting that the *Lm*Chr26 portion does not contain the telomeric end of the chromosome. The *T. cruzi* portion containing the corresponding region on chromosome 8 is also mostly composed of homologs (Figs. 1a and 2a) as are the two larger *T. cruzi* assemblies represented in the alignments. These scaffolds do not overlap but represent the adjoining homologous regions on *Tb*ChrIX. They also resemble the subtelomeric regions of *Leishmania* chromosomes 1 and 15 as represented in Fig. 3.

In a similar manner, we compared a portion of *T. brucei* chromosome I (*Tb*ChrI; [9]) to its homologous regions in *T. cruzi* and *L. major* (Fig. 1b). The *T. cruzi* 100 kb contig is comprised largely of a previously published segment from chromosome 3 [14]. To the right of the strand-switch region (which corresponds to the presence of retrotransposon-like elements; Fig. 1b, black arrows) on *Tb*chrI, we observe a

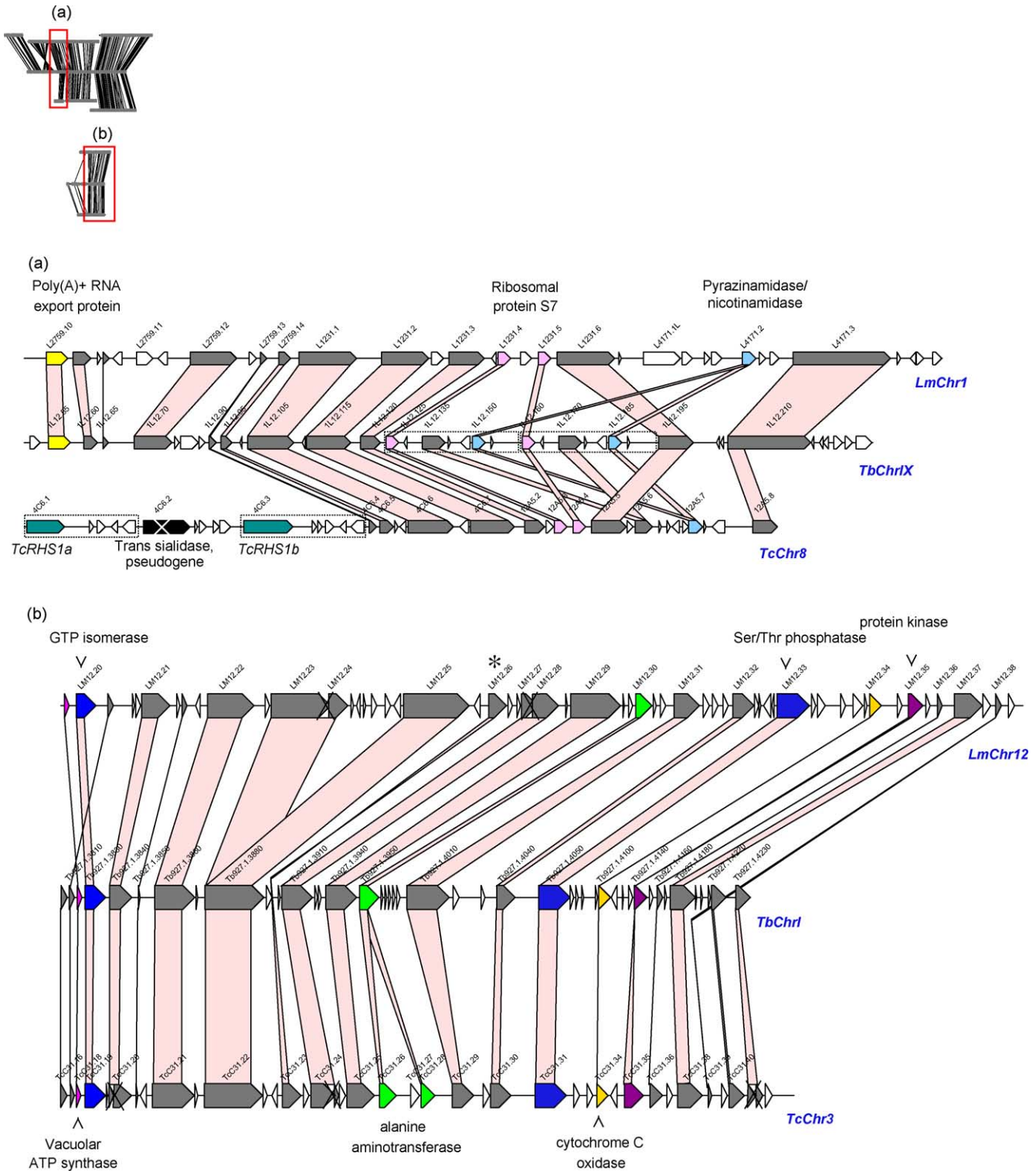

Fig. 2. Detailed representation of regions of three-way synteny in *L. major*, *T. brucei* and *T. cruzi*. Red rectangles delineate zoom-in areas from Fig. 1. Grey boxes correspond to 'conserved hypothetical proteins' which are proteins with homologs in other species but with no assigned functions. Colored boxes refer to genes with assigned functions; white boxes are putative CDSs ('hypothetical proteins') with no assigned function. The 'X' marks indicate pseudogenes. Locus names appearing above genes correspond to their systematic names as submitted to GenBank and assigned in the TIGR database (http://www.tigr.org/tdb/e2k1/tba1/LocusNameSearch.shtml). Dashed boxes correspond to segmental duplications. (a) Three-way alignment of *Lm*Chr1, *Tb*ChrIX and *Tc*Chr8 segments. (b) Three-way alignment of *Lm*Chr12, *Tb*ChrI and *Tc*Chr3 segments. The *Lm*Chr12 gene names are as they appear in GenBank accession number AL390114. These names are not stable because temporary names are assigned for whole chromosome shotgun sequence before finished contiguous chromosomes are assembled. Locus names appearing above genes for *Tb*ChrI correspond to their systematic names assigned in GeneDB (http://www.genedb.org).

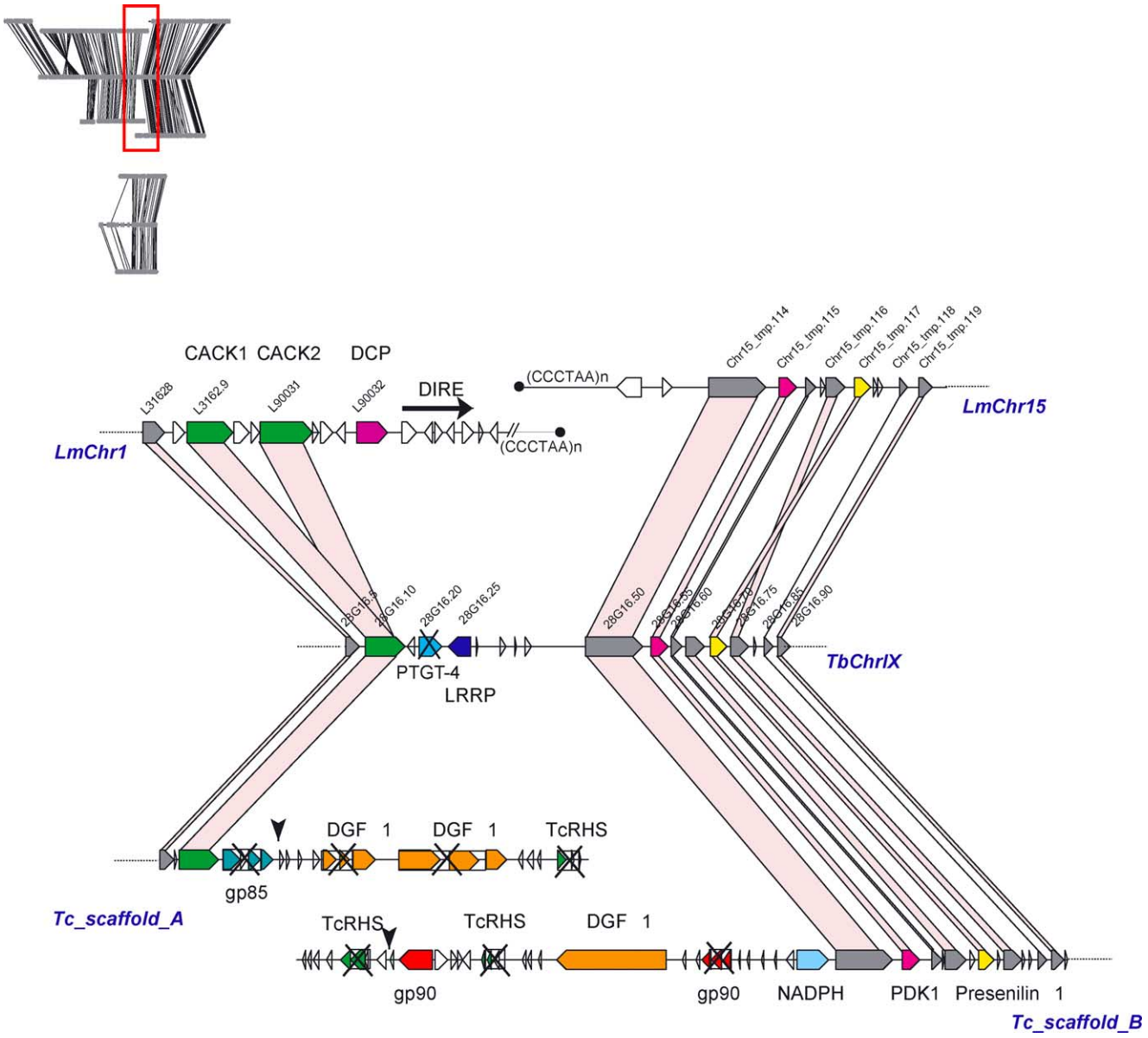

Fig. 3. Detailed representation of the region of synteny where *T. brucei* chromosome IX overlaps the ends of *L. major* chromosome 1 and 15 and two *T. cruzi* scaffolds. Red rectangle delineates zoom-in area from Fig. 1. Grey boxes correspond to 'conserved hypothetical proteins' which are proteins with homologs in other species but no assigned functions; colored boxes refer to genes with assigned functions; white boxes are putative CDSs ('hypothetical proteins') with no assigned functions. The 'X' marks indicate pseudogenes. The lines that appear at the telomeric end of *Lm*Chr1 indicate that there are >20 kb of subtelomeric repeats [10]. The full black circles represent telomeric ends. The black arrow represents the location of the *Lm*Chr1-DIRE. The two vertical black arrow heads indicate the location of the retroposon-like SIRE sequences. Locus names appearing above genes correspond to their systematic names as submitted to GenBank for *Lm*Chr1, or assigned in the TIGR database (www.tigr.org/tdb/e2k1/tba1/LocusNameSearch.shtml) for *Tb*ChrIX, and GeneDB (http://www.genedb.org) for *Lm*Chr15. The *Lm*Chr15 names are not stable because temporary names are assigned for whole chromosome shotgun sequence before finished contiguous chromosomes are assembled. CACK: calcium-activated potassium channel; DCP: dipeptidyl carboxypeptidase; PDK1: 3-phosphoinositide-dependent protein kinase-1; LRRP: leucine rich repeat protein; PTGT-4: putative glucose transporter. This transporter was previously found to be a single copy gene [45]. The copy of this gene present here is non-functional as the open reading frame is interrupted by 11 frame shifts. DGF-1: protein 1 of a *T. cruzi* dispersed gene family.

strong conservation of gene order in all three organisms with one apparent gene inversion in *L. major* as compared to the *Trypanosoma* (the first and third genes in the alignment in Fig. 2b). There is also one gene duplication in *Lm*Chr12 (with one copy being a pseudogene), as compared to both trypanosome species; and there is one gene duplication (alanine aminotransferase, green boxes) in *Tc*Chr3 as compared to both *Tb*ChrI and *Lm*Chr12. There is a gene found in *Lm*Chr12 and *Tc*Chr3 that apparently has been lost from *Tb*ChrI (see location of asterisk (∗) in Fig. 2b, LM12.26/TcC31.23). To the left of the strand-switch, only one gene is conserved in all three organisms (Fig. 1b). None

of the hypothetical genes appearing at the left-most end of the *Lm*Chr12 contig are homologs of *Tc*Chr3 or *Tb*ChrI genes while some level of synteny appears to be maintained between *Tc*Chr3 and *Tb*ChrI.

In *T. cruzi*, regions of synteny are interrupted by a set of genes belonging to a *RHS* multigene family (green arrows in Fig. 1a; details in Figs. 2a and 3). *RHS* genes, originally described in *T. brucei*, were found to contain hot spots for retroelement insertion [15,16] and appear to be primarily located near chromosomal telomeres [8,9]. It is not clear at this point whether *RHS* genes also exist at telomeric ends of *T. cruzi* chromosomes. Several *T. cruzi* subtelomeric regions, however, have been shown to contain copies of the gp90 and gp85/sialidase gene families as well as sequences derived from a retroposon-like element (SIRE) [17]. Based on their gene content, the two *T. cruzi* scaffolds displayed in the alignment to *Tb*ChrIX both contain ends that are reminescent of telomeric/subtelomeric ends (Fig. 3).

Of particular interest in all alignments (Fig. 2) is the conservation of hypothetical proteins in the three trypanosomatids. These hypothetical proteins do not match genes in any other organisms searched to date. Their preserved synteny in the trypanosomatids is an indication that they are true genes. Regions of breaks in synteny correspond generally to the presence of small hypothetical genes that are unlikely to all code for proteins (exemplified by the region on *Lm*Chr1 highlighted with a star in Fig. 1a) and to the insertion of larger unique genes (represented by the symbol psi (Ψ)) within a block of conserved genes. While these larger genes may not have an assigned function, they are of interest because of their presence in one particular genome.

An amino acid comparison in all three-way regions reveals an average of 40–45% identity between *L. major* and *Trypanosoma* homologs, whereas *T. brucei* and *T. cruzi* homologs have an average percent identity of more than 55%. These levels of protein conservation are consistent with the apparent ancient divergence of these lineages. Given this high level of sequence divergence, it is surprising to find such extensive synteny between these organisms, since there should have been wide opportunities for genome rearranging mutations. Comparisons of other eukaryotic genomes suggest that the level of synteny observed here is greater than would be expected for the age of the divergence of these species. For example, the nematodes *Caenorhabditis elegans* and *Caenorhabditis briggsae* are estimated to have diverged from each other approximately 25–50 Ma but their genomes nevertheless only align on average for less than 9 kb before an interruption in the conservation of their gene order [18]. While there is extensive synteny among closely related mammalian genomes [5], the mammalian groups compared are not as old as the trypanosomatids.

We propose that the high level of synteny even in the face of extensive sequence divergence is maintained by selection. The fact that the degree of synteny is about the same within genera (*T. cruzi* versus *T. brucei*) as compared to between genera (*L. major* with both trypanosomes) suggests that the selective forces have been operating since these lineages diverged. Why would gene order be so highly constrained in the trypanosomatids? In prokaryotes, gene order is frequently maintained for clusters of genes of related function in the form of operons [8,9]. Similarly, while 60% of the *C. elegans* genome is thought to be susceptible to rearrangements, the regions that appear stable correspond to transcriptionally co-regulated genes [18]. However, while there is one reported case in *T. cruzi*, where the enzymes involved in de novo pyrimidine biosynthesis are encoded by genes part of the same polycistronic transcription unit [19], overall the transcriptional units in the trypanosomatids do not appear to be grouped by function. The physical linkage of several isoforms of the same gene (clusters of repeated genes) as is the case for GP63, *trans*-sialidase and some transporters [20–23], would not explain the large regions of synteny observed. In some prokaryotes, high levels of synteny have been observed in species that are unable to carry out homologous recombination (thereby preventing genome rearrangements) [24]. This is unlikely in the trypanosomatids since the components for homologous recombination exist and this process has been exploited for genetic manipulations and knockouts. Furthermore, a homologue of the *Escherichia coli* recA was characterized in *Leishmania* [25] and genetic exchanges were found to occur extensively in *T. cruzi* [26].

One feature that sets the trypanosomatids apart from many other eukaryotes is the fact that trypanosomatid genes are transcribed as large polycistronic precursor RNAs that are subsequently cleaved into monocistronic mRNAs (reviewed in [27]). Gene expression is believed to be regulated in a post-transcriptional manner. While protein-coding genes are transcribed by a RNA polymerase II, no RNA polymerase II promoters have been clearly identified to date. The reason for the existence of these polycistronic messages has yet to be discerned. One possibility is that the origins of replication may be in the regions between polycistronic messages [28]. If so, the polycistronic messages may exist to allow for the replication and transcription to be co-directional. This has been shown to be an important factor in maintaining gene order and transcription direction in prokaryotes [24,29,30]. With such large polycistronic messages it seems even more likely that maintaining the co-direction of replication and transcription would be important. If this is true, this would lead to selection against most genome rearrangements, since most would disrupt the coordination of the directions.

The chromosomal rearrangements detected in the alignments presented here are associated with strand-switch regions or chromosome ends. There is one large segment inversion between *Lm*Chr1 and *Tb*ChrIX, for example, that encompasses the entire region between the “left” telomere and the strand-switch region of *Lm*ChrI (Fig. 1a). Interestingly, strand-switch areas appear to be correlated with sites of non-LTR retrotransposons as all retroelements observed in the alignments studied here are located in these particular regions (black arrows in Fig. 1). *T. brucei* chromosome II [8] has many retroelements present in areas

where there appears to be divergence of directional gene clusters. Because recombination between retrotransposons can influence genome organization, such events may have contributed to the generation of these strand-switch areas.

Non-LTR retrotransposons have been described in both *T. brucei* (*ingi*/RIME/SLACS) and *T. cruzi* (L1Tc/NARTc/CZAR), but no transposable elements of any kind have been observed so far in *Leishmania* species [15,31–35]. The autonomous *T. brucei ingi* (5.2 kb) and *T. cruzi* L1Tc (4.9 kb) retroelements are the most abundant elements described in these genomes. They encode 1657 and 1574 amino acid proteins, respectively, which share 23.7% identity [15]. The *ingi*- and L1Tc-encoded proteins contain four domains, i.e. endonuclease, reverse transcriptase, RNaseH and DNA-binding motifs [31,36,37], which are presumably involved in their own retrotransposition. All *ingi* retroelements sequenced to date have more than 90% nucleotide identity (unpublished results). The same was observed for L1Tc. This indicates these retroelements remain potentially active in the modern genome, or have been until recently, as functionality provides the selective pressure that slows down the rate of evolution.

As mentioned above, all the non-LTR retrotransposons or retrotransposon-like sequences identified in these analyzed regions appear to be located in areas of chromosome inversions, strand-switch and chromosome ends. This is exemplified by the presence of a truncated *ingi* in *Tb*ChrIX flanking a large inversion, relative to *Lm*ChrI (Fig. 1a). The presence of retroelements or retrotransposon-like elements was also noted in the homologous portions of *Tb*ChrI and *Tc*Chr3 which contain a strand-switch region of approximately 50 and 25 kb in length, respectively (black arrows in Fig. 1b). In the *Tb*ChrI portion, two full length *ingi*s are found in the proximity of a newly identified degenerated retrotransposon-like element that we call ‘DIRE’ (degenerated *Ingi*/L1Tc-related element). Two other DIREs are observed in *T. cruzi* in the homologous strand-switch region of *Tc*Chr3 (Fig. 1b). In addition, another retroelement-like sequence (*Lm*ChrI-DIRE) is found at one end of *L. major* chromosome 1 (Fig. 1a). Retroposon-like elements have not previously been observed in the *Leishmania* species. The location of a DIRE at the end of *Lm*Chr1, 2 kb upstream of the potential centromere [38], corresponds to the junction where homology on *Tb*ChrIX begins to *Lm*Chr15 (Figs. 1a and 3). The DIREs appear to be vestigial elements as they have conserved some domain homologies to retrotransposons such as *ingi*s and/or L1Tcs. Peptide sequence reconstructions of the DIREs from *T. brucei*, *T. cruzi* and *L. major* display amino acid identities to *ingi* ranging from 22 to 30%. This sequence divergence is a strong indication that the last retrotransposition event is not recent. This represents the first evidence that retroelements may have been active in the common ancestor of all three species. The location of a degenerate retrotransposon-like element at the end of *Lm*Chr1—i.e. at the junction where homology on *Tb*ChrIX to *Lm*Chr15 begins—as well as the presence of SIRE elements in the homologous *T. cruzi* regions, is a strong indication of the involvement of retrotransposon elements in chromosomal arrangements.

An interesting result of our analysis is the finding that the *Leishmania* chromosomes appear to be fragmented versions of the comparable chromosome from *T. brucei*. One complete *Leishmania* chromosome for example fits into one small portion of a much larger *T. brucei* chromosome. This is likely true for the whole genome since all trypanosomatids have genomes that are roughly of equivalent sizes, yet pulsed-field gel (PFG) analyses reveal that *T. brucei* has 11 chromosomal pairs while *T. cruzi* and *Leishmania* have 20–40 and 36, respectively, [10,39,40]. It is not completely understood how cleavage or fusion events through evolution may have shaped the current architecture of trypanosomatid genomes. The data reported here, however, is consistent with structural similarities observed with PFG between *Leishmania* and *T. cruzi* when compared to *T. brucei*. In both *Leishmania* and *T. cruzi*, the segments corresponding to the *Tb*ChrIX region detailed in Fig. 3 seem to come from two different chromosomes. In the case of *Leishmania*, the *Tb*ChrIX region unambiguously matches telomeric and subtelomeric regions of *Lm*Chr1 and *Lm*Chr15. The corresponding *T. cruzi* portions also appear to be the subtelomeric regions of the two *T. cruzi* chromosome homologs since they contain multiple copies of genes (gp85, gp90) and retroelements that have been described as landmarks of *T. cruzi* chromosome ends. We examined the DNA region between *Tb*ChrIX genes 28G16.25 and 28G16.50 for clues on whether the *T. brucei* chromosomes are the result of chromosome fusion to form the fewer and larger chromosomes, rather than the *T. cruzi* and *Leishmania* chromosomes being generated by chromosome cleavage. No particular sequence properties could be detected except for a slightly higher AT content (55%) in that portion when compared to the rest of *Tb*ChrIX (52%). Based on recent phylogenetic analyses of a diverse selection of trypanosomatid species that strongly suggest that the genus *Trypanosoma* is monophyletic [41–44], we propose a fusion scenario. If the ancestral chromosomal state was that observed in *T. brucei*, two independent fragmentation events would be required to explain the current chromosomal architecture, one in the lineage leading to *Leishmania* and another in the one leading to *T. cruzi*. If, alternatively, the ancestral state corresponded to smaller chromosomes, only one event has to have occurred, that is a chromosomal fusion in the lineage leading to *T. brucei*. The fact that the subtelomeric regions have similar locations in both *Leishmania* and *T. cruzi* further supports the more parsimonious fusion scenario, as otherwise two fragmentation events would have to occur independently in the same chromosomal location. This particular region does not seem to differ from the rest of the chromosome in such a way that would make it more prone to fragmentation.

With complete genomes available within the next year for whole genome comparative analyses, it will be interesting to determine where synteny is interrupted. The

question remains as to where species-specific genes will be located. Do they tend to be clustered in particular areas of the chromosomes (pathogenicity islands) or will they be found as indels (insertion/deletion) peppered throughout the genome? Examination of insertion/deletion events, gene duplication between species, and comparative chromosomal architecture, will allow even better insight into the evolutionary relationship of the trypanosomatids.

Acknowledgements

The work at the Institute for Genomic Research (TIGR), the Seattle Biomedical Research Institute (SBRI) and the Karolinska Institute was supported by grants from the National Institute for Allergy and Infectious Diseases, National Institutes of Health (NIAID/NIH) to N.E.S. (AI43062, AI45038), K.S. (AI49599, AI45039) and B.A. (AI45061). The work at the Sanger Institute was supported by grants from the Wellcome Trust. The authors would like to acknowledge John Donelson and Joanna DaSilva for editing the manuscript, Jinming Song and Samir Kaul for useful discussions, and John Swindle for donating the *T. cruzi* BAC 12A5.

References

[1] Baldauf SL, Roger AJ, Wenk-Siefert I, Doolittle WF. A kingdom-level phylogeny of eukaryotes based on combined protein data. Science 2000;290:972–7.
[2] Overath P, Haag J, Lischke A, O'hUigin C. The surface structure of trypanosomes in relation to their molecular phylogeny. Int J Parasitol 2001;31:468–71.
[3] Stevens JR, Gibson WC. The evolution of pathogenic trypanosomes. Cad Saude Publica 1999;15:673–84.
[4] Stevens JR, Noyes HA, Schofield CJ, Gibson W. The molecular evolution of Trypanosomatidae. Adv Parasitol 2001;48:1–56.
[5] O'Brien SJ, Menotti-Raymond M, Murphy WJ, Nash WG, Wienberg J, Stanyon R, et al. The promise of comparative genomics in mammals. Science 1999;286:458–81.
[6] Borst P, Ulbert S. Control of VSG gene expression sites. Mol Biochem Parasitol 2001;114:17–27.
[7] Myler PJ, Audleman L, deVos T, Hixson G, Kiser P, Lemley C, et al. *Leishmania major* Friedlin chromosome 1 has an unusual distribution of protein-coding genes. Proc Natl Acad Sci USA 1999;96:2902–6.
[8] El-Sayed NM, Ghedin E, Song J, MacLeod A, Bringaud F, Larkin C, et al. The sequence and analysis of *Trypanosoma brucei* chromosome II. Nucleic Acids Res 2003;31:4856–63.
[9] Hall N, Berriman M, Lennard N, Harris B, Hertz-Fowler C, Bart-Delabesse EN, et al. The DNA sequence of chromosome I of an African trypanosome: gene content, chromosome organisation, recombination and polymorphism. Nucleic Acids Res 2003;31:4864–73.
[10] Sunkin SM, Kiser P, Myler PJ, Stuart K. The size difference between *Leishmania major* friedlin chromosome one homologues is localized to sub-telomeric repeats at one chromosomal end. Mol Biochem Parasitol 2000;109:1–15.
[11] Henriksson J, Porcel B, Rydaker M, Ruiz A, Sabaj V, Galanti N, et al. Chromosome specific markers reveal conserved linkage groups in spite of extensive chromosomal size variation in *Trypanosoma cruzi*. Mol Biochem Parasitol 1995;73:63–74.
[12] Delcher AL, Harmon D, Kasif S, White O, Salzberg SL. Improved microbial gene identification with GLIMMER. Nucleic Acids Res 1999;27:4636–41.
[13] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol 1990;215:403–10.
[14] Andersson B, Aslund L, Tammi M, Tran AN, Hoheisel JD, Pettersson U. Complete sequence of a 93.4-kb contig from chromosome 3 of *Trypanosoma cruzi* containing a strand-switch region. Genome Res 1998;8:809–16.
[15] Bringaud F, Garcia-Perez JL, Heras SR, Ghedin E, El Sayed NM, Andersson B, et al. Identification of non-autonomous non-LTR retrotransposons in the genome of *Trypanosoma cruzi*. Mol Biochem Parasitol 2002;124:73–8.
[16] Bringaud F, Biteau N, Melville SA, Hez S, El-Sayed NM, Leech V, et al. A new, expressed multigene family containing a hot spot for insertion of retroelements is associated with polymorphic subtelomeric regions of *Trypanosoma brucei*. Eukaryotic Cell 2002;1:137–51.
[17] Chiurillo MA, Cano I, da Silveira JF, Ramirez JL. Organization of telomeric and sub-telomeric regions of chromosomes from the protozoan parasite *Trypanosoma cruzi*. Mol Biochem Parasitol 1999;100:173–83.
[18] Kent WJ, Zahler AM. Conservation, regulation, synteny, and introns in a large-scale *C. briggsae–C. elegans* genomic alignment. Genome Res 2000;10:1115–25.
[19] Gao G, Nara T, Nakajima-Shimada J, Aoki T. Novel organization and sequences of five genes encoding all six enzymes for de novo pyrimidine biosynthesis in *Trypanosoma cruzi*. J Mol Biol 1999;285:149–61.
[20] Bringaud F, Vedrenne C, Cuvillier A, Parzy D, Baltz D, Tetaud E, et al. Conserved organization of genes in trypanosomatids. Mol Biochem Parasitol 1998;94:249–64.
[21] Egima CM, Briones MR, Freitas Junior LH, Schenkman RP, Uemura H, Schenkman S. Organization of trans-sialidase genes in *Trypanosoma cruzi*. Mol Biochem Parasitol 1996;77:115–25.
[22] Bringaud F, Baltz T. African trypanosome glucose transporter genes: organization and evolution of a multigene family. Mol Biol Evol 1994;11:220–30.
[23] Steinkraus HB, Greer JM, Stephenson DC, Langer PJ. Sequence heterogeneity and polymorphic gene arrangements of the *Leishmania guyanensis* gp63 genes. Mol Biochem Parasitol 1993;62:173–85.
[24] Suyama M, Bork P. Evolution of prokaryotic gene order: genome rearrangements in closely related species. Trends Genet 2001;17:10–3.
[25] McKean PG, Keen JK, Smith DF, Benson FE. Identification and characterisation of a RAD51 gene from *Leishmania major*. Mol Biochem Parasitol 2001;115:209–16.
[26] Gaunt MW, Yeo M, Frame IA, Stothard JR, Carrasco HJ, Taylor MC, et al. Mechanism of genetic exchange in American trypanosomes. Nature 2003;421:936–9.
[27] Donelson JE, Gardner MJ, El Sayed NM. More surprises from Kinetoplastida. Proc Natl Acad Sci USA 1999;96:2579–81.
[28] McDonagh PD, Myler PJ, Stuart K. The unusual gene organization of *Leishmania major* chromosome 1 may reflect novel transcription processes. Nucleic Acids Res 2000;28:2800–3.
[29] Eisen JA, Heidelberg JF, White O, Salzberg S. Evidence for symmetric chromosomal inversions around the replication origin in bacteria. Genome Biol 2000;1:0011.1–9.
[30] Mackiewicz P, Mackiewicz D, Kowalczuk M, Cebrat S. Flip-flop around the origin and terminus of replication in prokaryotic genomes. Genome Biol 2001;2:1004.1–4.
[31] Martin F, Maranon C, Olivares M, Alonso C, Lopez MC. Characterization of a non-long terminal repeat retrotransposon cDNA (L1Tc) from *Trypanosoma cruzi*: homology of the first ORF with the ape family of DNA repair enzymes. J Mol Biol 1995;247:49–59.

[32] Villanueva MS, Williams SP, Beard CB, Richards FF, Aksoy S. A new member of a family of site-specific retrotransposons is present in the spliced leader RNA genes of *Trypanosoma cruzi*. Mol Cell Biol 1991;11:6139–48.
[33] Aksoy S, Williams S, Chang S, Richards FF. SLACS retrotransposon from *Trypanosoma brucei* gambiense is similar to mammalian LINEs. Nucleic Acids Res 1990;18:785–92.
[34] Kimmel BE, ole-MoiYoi OK, Young JR. Ingi, a 5.2-kb dispersed sequence element from *Trypanosoma brucei* that carries half of a smaller mobile element at either end and has homology with mammalian LINEs. Mol Cell Biol 1987;7:1465–75.
[35] Hasan G, Turner MJ, Cordingley JS. Complete nucleotide sequence of an unusual mobile element from *Trypanosoma brucei*. Cell 1984;37:333–41.
[36] Olivares M, Alonso C, Lopez MC. The open reading frame 1 of the L1Tc retrotransposon of *Trypanosoma cruzi* codes for a protein with apurinic-apyrimidinic nuclease activity. J Biol Chem 1997;272:25224–8.
[37] Pays E, Murphy NB. DNA-binding fingers encoded by a trypanosome retroposon. J Mol Biol 1987;197:147–8.
[38] Dubessay P, Ravel C, Bastien P, Stuart K, Blaineau C, Pages M. Mitotic stability of a coding DNA sequence-free version of *Leishmania major* chromosome 1 generated by targeted chromosome fragmentation. Gene 2002;289:151–9.
[39] Santos MR, Cano MI, Schijman A, Lorenzi H, Vazquez M, Levin MJ, et al. The *Trypanosoma cruzi* genome project: nuclear karyotype and gene mapping of clone CL Brener. Mem Inst Oswaldo Cruz 1997;92:821–8.
[40] Wincker P, Ravel C, Blaineau C, Pages M, Jauffret Y, Dedet JP, et al. The Leishmania genome comprises 36 chromosomes conserved across widely divergent human pathogenic species. Nucleic Acids Res 1996;24:1688–94.
[41] Haag J, O'hUigin C, Overath P. The molecular phylogeny of trypanosomes: evidence for an early divergence of the Salivaria. Mol Biochem Parasitol 1998;91:37–49.
[42] Lukes J, Jirku M, Dolezel D, Kral'ova I, Hollar L, Maslov DA. Analysis of ribosomal RNA genes suggests that trypanosomes are monophyletic. J Mol Evol 1997;44:521–7.
[43] Wright AD, Li S, Feng S, Martin DS, Lynn DH. Phylogenetic position of the kinetoplastids, *Cryptobia bullocki*, *Cryptobia catostomi*, and *Cryptobia salmositica* and monophyly of the genus Trypanosoma inferred from small subunit ribosomal RNA sequences. Mol Biochem Parasitol 1999;99:69–76.
[44] Stevens JR, Noyes HA, Dover GA, Gibson WC. The ancient and divergent origins of the human pathogenic trypanosomes, *Trypanosoma brucei* and *T. cruzi*. Parasitology 1999;118:107–16.
[45] Bayele HK, Eisenthal RS, Towner P. Complementation of a glucose transporter mutant of Schizosaccharomyces pombe by a novel *Trypanosoma brucei* gene. J Biol Chem 2000;275:14217–22.