RESEARCH ARTICLE

# The Genome of the African Trypanosome *Trypanosoma brucei*

Matthew Berriman,[1]* Elodie Ghedin,[2,3] Christiane Hertz-Fowler,[1] Gaëlle Blandin,[2] Hubert Renauld,[1] Daniella C. Bartholomeu,[2] Nicola J. Lennard,[1] Elisabet Caler,[2] Nancy E. Hamlin,[1] Brian Haas,[2] Ulrike Böhme,[1] Linda Hannick,[2] Martin A. Aslett,[1] Joshua Shallom,[2] Lucio Marcello,[4] Lihua Hou,[2] Bill Wickstead,[5] U. Cecilia M. Alsmark,[6] Claire Arrowsmith,[1] Rebecca J. Atkin,[1] Andrew J. Barron,[1] Frederic Bringaud,[7] Karen Brooks,[1] Mark Carrington,[8] Inna Cherevach,[1] Tracey-Jane Chillingworth,[1] Carol Churcher,[1] Louise N. Clark,[1] Craig H. Corton,[1] Ann Cronin,[1] Rob M. Davies,[1] Jonathon Doggett,[1] Appolinaire Djikeng,[2] Tamara Feldblyum,[2] Mark C. Field,[9] Audrey Fraser,[1] Ian Goodhead,[1] Zahra Hance,[1] David Harper,[1] Barbara R. Harris,[1] Heidi Hauser,[1] Jessica Hostetler,[2] Al Ivens,[1] Kay Jagels,[1] David Johnson,[1] Justin Johnson,[2] Kristine Jones,[2] Arnaud X. Kerhornou,[1] Hean Koo,[2] Natasha Larke,[1] Scott Landfear,[10] Christopher Larkin,[2] Vanessa Leech,[9] Alexandra Line,[1] Angela Lord,[1] Annette MacLeod,[4] Paul J. Mooney,[1] Sharon Moule,[1] David M. A. Martin,[11] Gareth W. Morgan,[12] Karen Mungall,[1] Halina Norbertczak,[1] Doug Ormond,[1] Grace Pai,[2] Chris S. Peacock,[1] Jeremy Peterson,[2] Michael A. Quail,[1] Ester Rabbinowitsch,[1] Marie-Adele Rajandream,[1] Chris Reitter,[9] Steven L. Salzberg,[2] Mandy Sanders,[1] Seth Schobel,[2] Sarah Sharp,[1] Mark Simmonds,[1] Anjana J. Simpson,[2] Luke Tallon,[2] C. Michael R. Turner,[13] Andrew Tait,[4] Adrian R. Tivey,[1] Susan Van Aken,[2] Danielle Walker,[1] David Wanless,[2] Shiliang Wang,[2] Brian White,[1] Owen White,[2] Sally Whitehead,[1] John Woodward,[1] Jennifer Wortman,[2] Mark D. Adams,[14] T. Martin Embley,[6] Keith Gull,[5] Elisabetta Ullu,[15] J. David Barry,[4] Alan H. Fairlamb,[11] Fred Opperdoes,[16] Barclay G. Barrell,[1] John E. Donelson,[17] Neil Hall,[1]† Claire M. Fraser,[2] Sara E. Melville,[9] Najib M. El-Sayed[2,3]*

African trypanosomes cause human sleeping sickness and livestock trypanosomiasis in sub-Saharan Africa. We present the sequence and analysis of the 11 megabase-sized chromosomes of *Trypanosoma brucei*. The 26-megabase genome contains 9068 predicted genes, including ~900 pseudogenes and ~1700 *T. brucei*–specific genes. Large subtelomeric arrays contain an archive of 806 variant surface glycoprotein (VSG) genes used by the parasite to evade the mammalian immune system. Most *VSG* genes are pseudogenes, which may be used to generate expressed mosaic genes by ectopic recombination. Comparisons of the cytoskeleton and endocytic trafficking systems with those of humans and other eukaryotic organisms reveal major differences. A comparison of metabolic pathways encoded by the genomes of *T. brucei*, *T. cruzi*, and *Leishmania major* reveals the least overall metabolic capability in *T. brucei* and the greatest in *L. major*. Horizontal transfer of genes of bacterial origin has contributed to some of the metabolic differences in these parasites, and a number of novel potential drug targets have been identified.

Human African trypanosomiasis, or sleeping sickness, primarily affects the poorest rural populations in some of the least developed countries of Central Africa (*1*). The incidence may approach 300,000 to 500,000 cases per year, and it is invariably fatal if untreated. The disease is caused by *Trypanosoma brucei*, an extracellular eukaryotic flagellate parasite, and current treatments are inadequate: Drugs for late-stage disease are highly toxic; there is no prophylactic chemotherapy and little or no prospect of a vaccine.

Livestock trypanosomiasis is caused by closely related *Trypanosoma* species. It has the greatest impact in sub-Saharan Africa, where the tsetse fly vector is common, and also occurs in Asia and South America.

We present the sequence and analysis of the 11 megabase-sized chromosomes of the *T. brucei* genome (Table 1). The nuclear genome also contains an unspecified number of small and intermediate-sized chromosomes (30 to 700 kb) (*2*, *3*) known to encode sequences similar to subtelomeric regions of megabase-sized chromosomes (*4*).

**Genome structure and content.** The genome data herein represent a haploid mosaic (*5*) of the diploid chromosomes and were determined by both whole chromosome shotgun (chromosomes 1 and 9 to 11) and bacterial artificial chromosome walking strategies (chromosomes 2 to 8). Most of the assembled chromosome sequences extend into the subtelomeric regions (the sequence between the telomere and the first housekeeping gene) (Plate 3 and table S1).

As previously observed on a smaller scale (*6*, *7*), both strands of the megabase chromosomes contain long, nonoverlapping gene clusters (Plate 3) that are probably transcribed as polycistrons and subsequently trans-spliced and polyadenylated. The unusual mechanisms of transcription and their implications for the parasite are discussed elsewhere (*8*). Just over 20% of the genome encodes subtelomeric genes

[1]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton CB10 1SA, UK. [2]The Institute for Genomic Research, Rockville, MD 20850, USA. [3]Department of Microbiology and Tropical Medicine, George Washington University, Washington, DC 20052, USA. [4]Wellcome Centre for Molecular Parasitology, University of Glasgow, 56 Dumbarton Road, Glasgow G11 6NU, UK. [5]Sir William Dunn School of Pathology, University of Oxford, South Parks Road, Oxford OX1 3RE, UK. [6]School of Biology, Devonshire Building, University of Newcastle upon Tyne, Newcastle NE1 7RU, UK. [7]Laboratoire de Génomique Fonctionnelle des Trypanosomatides, Université Victor Segalen Bordeaux II, UMR-5162 CNRS, 33076 Bordeaux cedex, France. [8]Department of Biochemistry, University of Cambridge, Cambridge CB2 1GA, UK. [9]Department of Pathology, University of Cambridge, Cambridge CB2 1QP, UK. [10]Oregon Health and Science University, 3181 SW Sam Jackson Park Road, Mail Code L474, Portland, OR 97239–3098, USA. [11]School of Life Sciences, Wellcome Trust Biocentre, University of Dundee, Dundee DD1 5EH, UK. [12]Department of Biological Sciences, Imperial College, London SW7 2AY, UK. [13]Institute of Biomedical and Life Sciences, Joseph Black Building, University of Glasgow, Glasgow G12 8QQ, UK. [14]Department of Genetics, Case Western Reserve University, 10900 Euclid Avenue, Cleveland, OH 44106, USA. [15]Department of Internal Medicine, Yale University School of Medicine, 333 Cedar Street, Post Office Box 208022, New Haven, CT 06520–8022, USA. [16]Christian de Duve Institute of Cellular Pathology and Catholic University of Louvain, Avenue Hippocrate 74-75, B-1200 Brussels, Belgium. [17]Department of Biochemistry, University of Iowa, Iowa City, IA 52242, USA.

*To whom correspondence should be addressed. E-mail: mb4@sanger.ac.uk (M.B.); nelsayed@tigr.org (N.M.E.-S.)

†Present address: Institute for Genomic Research, Rockville, MD 20850, USA.

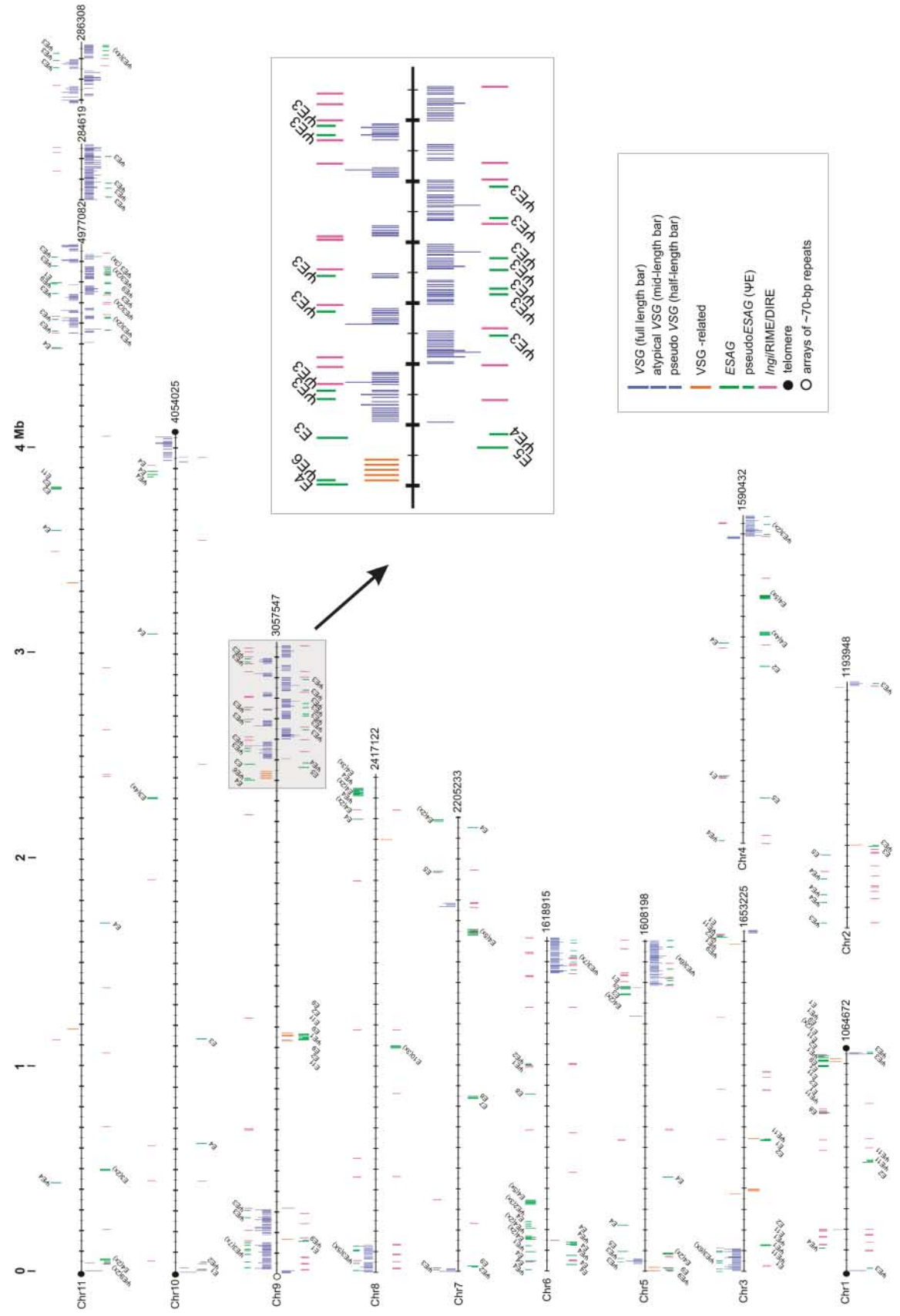


**Fig. 1.** Distribution of *VSG*s and *ESAG*s on chromosomes 1 to 11 of *T. brucei*. The four different *VSG* categories are defined in the text: *VSG*, atypical *VSG*, *VSG* pseudogene (gene fragments and full-length pseudogenes), and *VSG*-related. *ESAG* family members are identified (e.g., E1 is *ESAG1*), and degenerate, frameshifted, and truncated *ESAG*s are represented by ψ. Solid circles indicate arrays of telomeric-repeat hexamers and/or known terminal subtelomeric repeats, whereas open circles depict arrays of ~70-bp repeats. The *VSG* and *ESAG* array at the right-hand end of chromosome 9 is highlighted as an example.

(Plate 3) (*9*), the majority of which are *T. brucei*–specific and relate to the parasite's capacity to undergo antigenic variation in the bloodstream of its mammalian host. Expansion of gene families by tandem duplication is a mechanism by which the parasites can increase expression levels to compensate for a general lack of transcriptional control (*8*). Analysis of protein families (table S2) gives a measure of the resources that the parasite commits to key cellular processes. Kinesins and kinesin-like proteins, protein kinases (*8*), and adenylate cyclases are among the largest families, and two as-yet uncharacterized families of more than 10 members each appear to be expanded in the *T. brucei* genome relative to *L. major* and *T. cruzi* genomes.

**Antigenic variation.** Antigenic variation enables evasion of acquired immunity during infection or in immune host populations. Invariant antigens on the surface of *T. brucei* are shielded by a dense coat of $10^7$ copies of a single variant surface glycoprotein (VSG) that must be replaced as antibodies against it arise (*10*, *11*). VSGs differ significantly in the hypervariable N-terminal domain, which is exposed to the immune system, whereas the C-terminal domain is more conserved and is buried in the coat. The genome was previously known to contain a large archive of an estimated 1000 nonexpressed *VSG* genes (*VSG*s) that are activated clonally during infection at a rate of up to one activation event per 100 cell doublings (*12*, *13*). These *VSG*s have a basic cassette organization, with one or more ~70–base pair (bp) repeats at their 5′ flank and sequence homology at their 3′ flank that includes parts of the coding sequence and the 3′ untranslated region (*14*). *VSG* activation relies on residence in specific transcription units known as bloodstream-form expression sites (BESs), which are immediately subtelomeric on megabase and intermediate chromosomes and possess additional co-transcribed expression site–associated genes (*ESAG*s). *VSG* switching normally involves duplication of *VSG* cassettes into BESs.

The first new insight of the genome analysis is that most sequenced silent *VSG*s are defective. It is not known what proportion of *VSG*s are intact among the intermediate chromosomes, minichromosomes, or the yet-to-be-sequenced subtelomeric regions. Of 806 analyzed (Fig. 1 and table S3), only 57 (7%) are fully functional (that is, encode all recognizable features of known functional VSGs), whereas 9% are atypical (complete genes possibly encoding proteins with inconsistent VSG folding or post-translational modification), 66% are full-length pseudogenes (with frameshifts and/or in-frame stop codons), and 18% are gene fragments, most of which encode C-terminal domains. This *VSG* archive could be a repository of defunct material, but it might also be an information pool of an unprecedented size among parasites. Antigen pseudogenes are known to exist in several bacterial and other protozoan pathogens and to contribute to immune evasion by partial gene conversions to become expressed mosaic genes (*15*). This combinatorial use of information economizes on the genome and, theoretically, can greatly increase the potential for variation. In African trypanosomes, it has long been known that VSGs of antigenic variants can be encoded by mosaics of pseudogenes and also by hybrids, where the N- and C-terminal domains are derived from separate genes (*16*, *17*). It now appears that more than one-third of the N-terminal domains encoded by full-length VSG are intact, whereas C-terminal domains are more degenerate, further indicating the likely importance of hybrid gene formation. It is also possible that a trypanosome may be able to reinfect a reservoir host previously exposed to trypanosomes by using the *VSG* archive to embark stochastically on a different pathway of mosaic formation.

The second new insight is that almost all *VSG*s form arrays, numbering 3 to ~250 (pseudo)genes, and that most of these are subtelomeric (Fig. 1). In contrast to genes in telomeric BESs, the majority of *VSG*s are oriented away from the telomere. This is consistent with other protozoa [except *Giardia* (*18*)], where antigen gene archives are at least partly subtelomeric, probably due to the ability of these chromosome domains to recombine with each other ectopically rather than merely between chromosome homologs (*19*). More than 90% of the *VSG*s have one or more ~70-bp repeats upstream. In addition, the non–long terminal repeat retransposons *ingi* and RIME are associated with the *VSG* arrays, in particular where coding strand switches occur, in keeping with the role of such elements in genome restructuring (*20*). We found no obvious organization of the most similar *VSG*s into subfamilies within arrays; *VSG*s are scattered apparently at random within and among arrays (fig. S1) or between homologous chromosomes (fig. S2).

VSGs can be subdivided into N- and C-terminal domains as defined by cysteine distribution and sequence homology (*21*). Two previously unknown types of C-terminal domain are encoded in the genome, and diversification among the three known N-terminal domains has occurred. Also, we identified a previously unknown set of 29 VSG-related genes that cluster distinctly from other VSGs in a multiple alignment and lack upstream ~70-bp repeats. The majority of these putative proteins do not contain cysteine residues in their C-terminal regions. They could have evolved novel functions, as has happened with the serum resistance-associated protein (*22*). This arrangement of genes resembles that of the VAR family in the malaria parasite *Plasmodium falciparum*; VARs encode variant erythrocyte surface proteins and most are subtelomeric, but some are internally located and differ from their subtelomeric counterparts (*23*).

Members of up to 11 families of *ESAG*s are normally associated with *VSG*s in the polycistronic BESs. All BESs appear to harbour an *ESAG6* and *ESAG7*, and most have a total of five to 10 ESAGs and pseudo-ESAGs (*24*). Of the 11 known ESAG families (Fig. 1 and table S4), *ESAG3* and *ESAG4* are remarkably abundant, but *ESAG3*s are preferentially associated with *VSG* arrays, where it is present exclusively as pseudogenes. A single copy of the tandem *ESAG6* and *ESAG7* (normally found at the BESs) resides outside BESs but not in the *VSG* arrays; it appears in tandem on chromosome 7 at an interruption in the conservation of synteny found between the Tritryps (*9*).

**Intracellular protein vesicle trafficking.** Membrane transport functions are critically important for the Tritryps, both in terms of their interactions with insect vectors and mammalian hosts as well as the recycling of surface membrane components, such as VSGs. The Tritryps have a polarized endomembrane system and restrict both exocytosis and endocytosis, which occur exclusively via a clathrin-dependent mechanism, to the flagellar pocket (*25*, *26*).

*T. brucei* has served as a paradigm for the study of transport processes in kinetoplastids, elucidating the role of cytoplasmic coat proteins (tables S5 and S6) and suggesting a fundamental difference in the mechanisms of

**Table 1.** Summary of the *T. brucei* genome. Genome size and chromosome numbers exclude intermediate and mini-chromosomes. Details of contig coverage for each chromosome are described in table S1. Intergenic regions are regions between protein-coding sequences (CDSs). The exact number of spliced leader (sl) RNA copies cannot be resolved in the assembly.

| Parameter | Number |
|---|---|
| *The genome* | |
| Size (bp) | 26,075,396 |
| G+C content (%) | 46.4 |
| Chromosomes | 11 |
| Sequence contigs | 30 |
| Percent coding | 50.5 |
| *Protein-coding genes* | |
| Genes | 9068 |
| Pseudogenes | 904 |
| Mean CDS length (bp) | 1592 |
| Median CDS length (bp) | 1242 |
| G+C content (%) | 50.9 |
| Gene density (genes per Mb) | 317 |
| *Intergenic regions* | |
| Mean length (bp) | 1279 |
| G+C content (%) | 41 |
| *RNA genes* | |
| transfer RNA | 65 |
| ribosomal RNA | 56 |
| slRNA | >28 |
| small nuclear RNA | 5 |
| small nucleolar RNA | 353 |

endocytosis between the Tritryps (*26*). Analysis of the *T. brucei* Rab guanine triphosphatases (GTPases), regulators of membrane transport, indicates a complex, highly regulated vesicular transport system (*27*). This complexity now appears to be mirrored in the *L. major* and *T. cruzi* genomes (tables S5 and S7). A second GTPase family includes adenosine diphosphate ribosylation factors (ARFs), ARF-like proteins (ARLs), and Sar1, regulators of membrane traffic and the cytoskeleton (table S5). Phylogenetic reconstruction of the divergent Tritryp ARF and ARL family indicated previously unknown isoforms, some of which are not found in metazoan organisms (fig. S3) (*28*). This likely acquisition of a large family of trypanosomatid-specific GTPases may reflect the extreme emphasis of the trypanosomatid cytoskeleton on tubulin- rather than actin-mediated mechanisms.

**Cytoskeleton.** Generally, the eukaryotic cytoskeleton is divided into three filament classes: microtubules, intermediate filaments, and actin microfilaments. Proteins of the intermediate filament network are very diverse and poorly conserved in evolution. It appears that no homologs to known intermediate filament genes (table S8) are encoded in the Tritryp genome sequences. Homologs of components of actin- and tubulin-based cytoskeleton, however, are readily identifiable (Fig. 2 and table S8). Compared to other eukaryotic organisms, the Tritryps appear to have a reduced dependence on the acto-myosin network balanced by an elaboration of the tubulin-based cytoskeleton.

Six members of the tubulin superfamily are present in the Tritryps: α-, β,- γ-, δ-, ε-, and ζ-tubulin. The presence of δ- and ε-tubulin is characteristic of organisms with basal bodies and flagella. Five putative genes encoding centrins, elongation factor (EF)–hand proteins functioning at microtubule-organizing centers (MTOCs), occur in each genome. Yeasts, in contrast, have a single MTOC and a single centrin gene. The existence of a family of centrins in the Tritryps may reflect the diversity of dispersed MTOCs involved in cytoskeleton organization in these parasites. Outside of the centrins, few homologs to the components of the mammalian centrosome or yeast spindle-pole body were identified in the three genomes (Fig. 2).

The Tritryp dependence on diverse microtubule function is best exemplified by the kinesin superfamily. Each genome encodes >40 putative kinesins (excluding near-identical copies) compared with only 31 in humans. Furthermore, these kinesins have a huge diversity of primary sequences (unlike, for example, the numerous kinesins of *Arabidopsis*). Some of these kinesins fall into previously identified functional clades, but many belong to novel clades. There are no homologs for the kinetochore motor, CENP-E, or the spindle motor, BimC. However, there has been an expansion and diversification of the mitotic centromere-associated kinesin family. Characteristic trilaminar kinetochore-like plaques are seen in kinetoplastid mitotic nuclei attached to the spindle (*29*). Generally, the proteins of the mitotic kinetochore are only poorly conserved despite the near-ubiquity of the structure itself (Fig. 3). However, this appears to be particularly pronounced in the case of the Tritryps, where there are no homologs of CENP-C/Mif2,

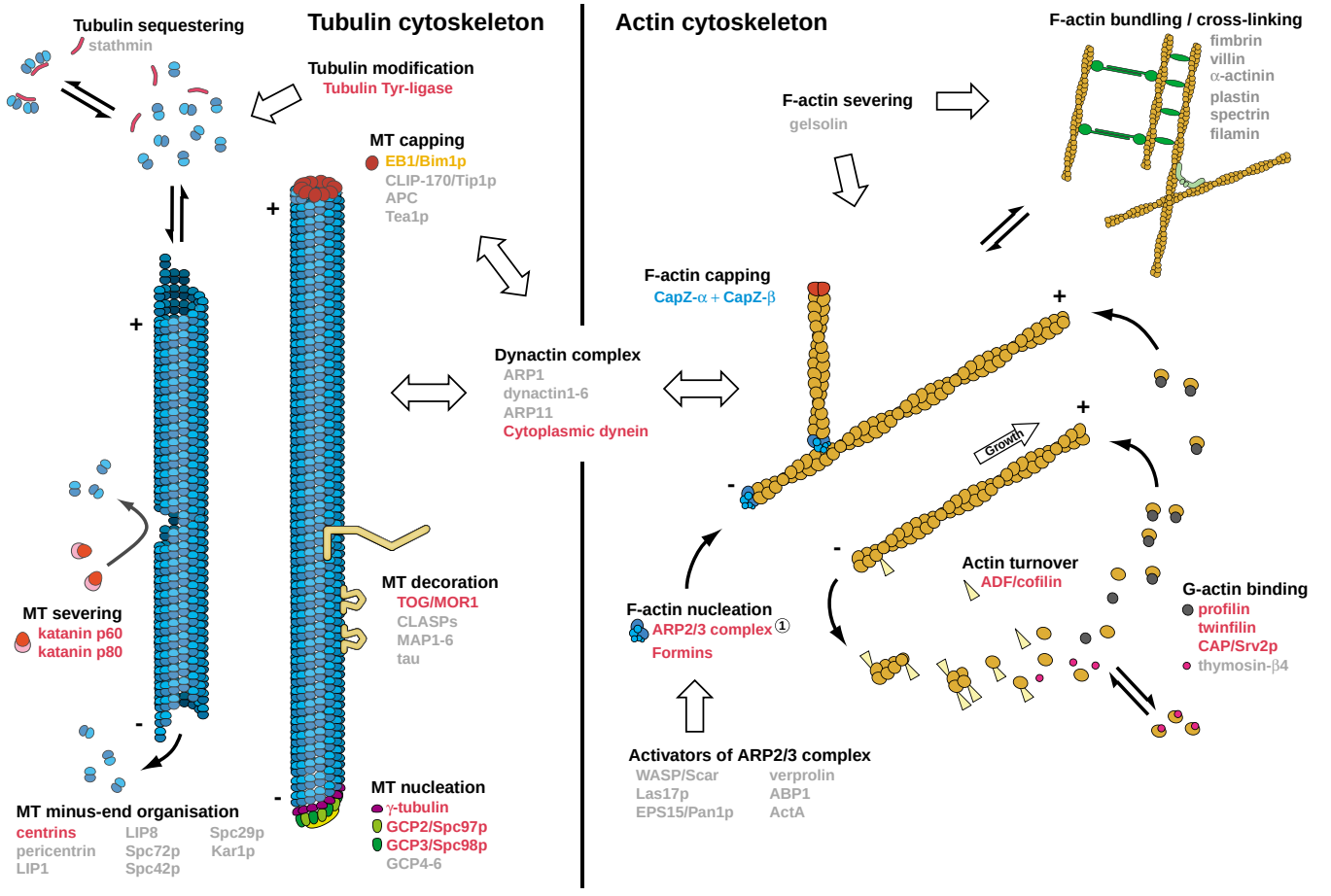


**Fig. 2.** Core cytoskeletal components in *T. brucei*, *T. cruzi*, and *L. major*. A schematic representation of the tubulin- and actin-based cytoskeleton. **(Left)** α- and β-tubulin subunits are depicted as light and dark blue circles, either assembled into the microtubule lattice or as dimers. **(Right)** Actin filaments composed of actin momomers (yellow circles). Black text indicates processes involved in cytoskeleton organization. Components for which the Tritryp genomes encode one or more homologs are shown in red text. Those that do not appear to be encoded by the three genomes are indicated in gray text. The CapZ proteins only present in *T. cruzi* are highlighted in blue text, whereas orange text points to homologs only encoded by the *T. brucei* and *T. cruzi* genomes. The Arp2/3 complex is divergent in *L. major*.

HEC1/Ndc80, or the centromeric checkpoint protein BUB1, each of which is highly conserved among yeast, plants, and mammals (Fig. 3). Indeed, it appears that the sole variant histone H3 in these genomes does not even function at the centromere (*30*).

The proteins of the flagellar axoneme appeared to be extremely well conserved. With the exception of tektin, there are homologs in the three genomes for all previously identified structural components as well as a full complement of flagellar motors and both complex A and complex B of the intraflagellar transport system (Fig. 4). Thus, the 9+2 axoneme, which arose very early in eukaryotic evolution, appears to be constructed around a core set of proteins that are conserved in organisms possessing flagella and cilia. The major divergence apparent in flagellar organization in Kinetoplastida is the augmentation of the axoneme by a paracrystalline extra-axonemal structure, the paraflagellar rod, that is unique to the Kinetoplastida and Euglenida.

No actin-based motility has been described in trypanosomatids; they do not exhibit the ruffles or pseudopodia seen in amoebae and metazoa or the myosin-based gliding motility of the Apicomplexa. Acto-myosin may not even be necessary for cytokinesis in these organisms. Actin is, however, involved in endocytosis (*31*). The genome sequences revealed putative homologs for proteins involved in monomeric-actin binding and actin filament nucleation (Fig. 2 and table S8) and for two myosin families (*32*). Surprisingly, the cohort of highly conserved proteins involved in severing, bundling, cross-linking, and capping of filamentous actin are missing [the latter is not the case for *T. cruzi* (*32*)]. Moreover, the dynactin complex is absent, indicating a major loss in the ability for cross-talk between the actin- and tubulin-filament networks (Fig. 2).

**Metabolism and transport.** Understanding the parasite's metabolism will underpin new drug discovery. To date, biochemical studies of trypanosomatids have been restricted to specific life-cycle stages that can be readily cultured in vitro or in vivo. In contrast, genome analysis provides a global view of the parasite's metabolic potential (Plate 2). In terms of overall capabilities, *T. brucei* has the most restrictive metabolic repertoire, followed by *T. cruzi*. This may reflect that, unlike *T. cruzi* or *L. major*, the parasite does not have an intracellular life cycle stage and therefore has greater access to nutrients in the plasma. Horizontal gene transfer may have provided additional metabolic versatility for the parasites. In almost 50 cases, there was strong evidence of putative horizontal transfer from bacteria into the Tritryp lineage (*5*) (table S11). Moreover, these provide us with candidate enzymes for new drug intervention points.

Across the three species, 633 genes (table S9) are annotated as transporters or channels, and 8113 are annotated as enzymes. Of the latter, more than 2000 have been classified with Enzyme Commission (EC) numbers (table S10).

**Carbohydrate metabolism.** The Tritryps possess a full complement of candidate genes necessary for uptake and degradation of glucose via glycolysis or the pentose phosphate shunt and the tricarboxylic acid (TCA) cycle (Plate 2). A limited number of hexose transporters are present in all species, but only *T. cruzi* possesses hexose phosphate transporters as seen in bacteria. This may reflect the fact that only *T. cruzi* amastigotes reside in the cytosol of the host cell with ready access to sugar phosphates. In contrast, only *L. major* appears capable of hydrolyzing disaccharides. This is consistent with the biology of the insect vectors: Tsetse and triatomines are obligate blood feeders, whereas sandflies also feed on nectar and aphid honeydew. *T. brucei* lacks several sugar kinases present in *L. major* or *T. cruzi*, possibly an adaptation to the restricted range of sugars available in plasma and in tsetse blood meals.

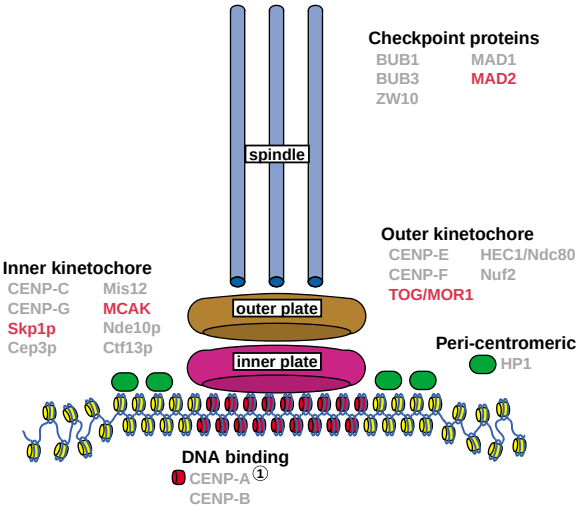


**Fig. 3.** Conservation of proteins of the kinetochore in *T. brucei*, *T. cruzi*, and *L. major*. A schematic representation of conserved components of a typical kinetochore is shown. The pair of plate-like kinetochores are present on the surface of the chromosome in the centromeric region, with the outer plate acting as microtubule attachment points radiating from spindle pole bodies. Kinetochore components for which the Tritryps encode one or more homologs are indicated in red text. Those that do not appear to be encoded by the three genomes are indicated in gray text. The kinetoplastid histone H3 variant does not appear to be centromeric (*30*).

Despite having potential pathways for complete oxidation, the Tritryps partially degrade glucose to succinate, ethanol, acetate, alanine, pyruvate, and glycerol (*33*) (Plate 2). L-lactate is not an end product because lactate dehydrogenase is absent from all three species, although D-lactate can be formed from methylglyoxal in *L. major*. A major route for formation of succinate from glucose in the insect stages is via an unusual nicotinamide adenine dinucleotide (NADH)–dependent fumarate reductase (*34*). In mitochondria, pyruvate is principally converted to acetate via an acetate:succinate coenzyme A (CoA)–transferase (*35*).

In many insects, including tsetse and sandflies, proline is a highly abundant energy source used in flight muscles. Proline is also a preferred energy source for the insect stages of the parasites. However, proline is only partially degraded to succinate (procyclic stage, *T. brucei*) or alanine, succinate, and other TCA intermediates (promastigote stage, *L. major*). In the Tritryps, neither an isocitrate lyase nor malate synthase could be identified, making a glyoxalate bypass in the TCA cycle unlikely.

**Electron transport and oxidative phosphorylation.** A fully functional mitochondrial electron transport system and adenosine triphosphate (ATP) synthase is present in the Tritryps. Several ubiquinone-linked dehydrogenases are present, including a flavin adenine dinucleotide (FAD)–dependent glycerol-3-phosphate dehydrogenase, which in *T. brucei* alone is linked to an alternative oxidase that represents a potential drug target (*36*).

In many organisms, heme is synthesized from glycine and succinyl-CoA and incorporated into hemoproteins, such as cytochromes. However, all trypanosomatids (except those with bacterial endosymbionts such as *Crithidia oncopelti*) are heme auxotrophs. Thus, the absence of the first two enzymes in heme biosynthesis (aminolevulinate synthase and aminolevulinate dehydratase) is not unexpected. It was surprising, however, to discover that genes for the last three enzymes (coproporphyrinogen III oxidase, protoporphyrinogen IX oxidase, and ferrochelatase) of heme biosynthesis are apparently present in *L. major*. Moreover, there is strong evidence (table S11 and fig. S4, A to C) that these genes may have been horizontally acquired from bacteria.

**Glycosylphosphatidylinositol anchor biosynthesis.** The surface of trypanosomatid plasma membranes contains a variety of glycoproteins (e.g., VSGs), phosphosaccharides (e.g., lipophosphoglycan), and glycolipids that are involved in immune evasion, attachment, or invasion. The most abundant of these are attached via glycosylphosphatidylinositol (GPI) anchors. GPI anchors are essential for parasite survival and are therefore a potential drug target (*37*). The GPI backbone contains mannose, glucosamine, and *myo*-inositol-phospholipid,

plus additional sugar modifications such as galactose (Plate 2). Pathways to the appropriate sugar nucleotide precursors are complete except for phosphoglucomutase, which was not identified in *T. brucei* despite ample biochemical evidence for its presence. Moreover, uridine diphosphate (UDP)–galactose can only be formed from glucose-6-phosphate via UDP-glucose (*38*), consistent with the absence of genes for conversion of galactose to this sugar in *T. brucei*. The *T. brucei* genome encodes a large family of UDP-galactose– or UDP-*N*-acetylglucosomine–dependent glycosyltransferases (table S2), nine of which appear to be pseudogenes associated with arrays of *VSG* pseudogenes. *Myo*-inositol can either be synthesized de novo from glucose-6-phosphate or salvaged via an inositol transporter (*39*) for the synthesis of phosphatidyl inositol, the first step in the GPI biosynthetic pathway.

Surface glycoproteins are attached to their GPI anchors via ethanolamine phosphate. In *T. cruzi* mucins, this linkage can optionally be replaced by aminoethylphosphonate (AEP). Enzymes for the complete synthesis of AEP from phosphoenol pyruvate are present only in *T. cruzi* (Plate 2) and could represent novel drug targets because they are absent from humans. The third (and last) step of the pathway (AEP transaminase) is also found in *L. major* and *T. cruzi*, and strong tree-based evidence supports its presence via horizontal acquisition from bacteria (table S11 and fig. S5). The second step of the pathway (phosphonopyruvate decarboxylase) is found in all three parasites, and homologs in other eukaryotes could not be found.

**Amino acid metabolism.** The Tritryps have some striking differences in amino acid metabolism (Plate 2, boxed). Amino acid transporters constitute one of the largest families of permeases in these parasites, with 29 predicted members in *L. major*, 38 in *T. brucei*, and 42 in *T. cruzi* (table S9). This is consistent with the Tritryps lacking biosynthetic pathways for the essential amino acids of humans and requiring exogenous proline as an energy source (except bloodstream *T. brucei*), glutamine for several biosynthetic pathways, cysteine as an additional sulfur source, and tyrosine for protein synthesis.

Most of the enzymes of the classical pathways for aromatic amino acid oxidation are missing. A putative biopterin-dependent phenylalanine-4-hydroxylase capable of converting phenylalanine to tyrosine could only be identified in *L. major*. However, *Leishmania* spp. are auxotrophic for tyrosine (*40*), so this gene (LmjF28.1280) may have another function. Candidate genes for transamination and reduction to the corresponding aromatic lactate derivative have been identified in all species. The role of this pathway is unclear, although secretion of these aromatic acids is found in *T. brucei* infections (*41*) and their presence in the central nervous system may result in the neurobehavioral disturbances (*42*) typically associated with human sleeping sickness.

The branched-chain amino acids can be converted to acyl-CoA derivatives in mitochondrion. However, no branched chain aminotransferase catalyzing the first step could be identified in *T. cruzi*. Leucine is converted to hydroxymethylglutaryl-CoA (HMG-CoA), which then can be incorporated directly into sterols via the isoprenoid synthetic pathway (*43*). Alternatively, HMG-CoA can be cleaved into acetyl-CoA and acetoacetate in *T. brucei* and *T. cruzi*, but the lyase is apparently absent from *L. major*. Isoleucine and valine use a similar pathway to form propionyl-CoA. However, the mitochondrial enzymes for further metabolism of propionyl-CoA to succinyl-CoA seem to be absent from both *T. brucei* and *T. cruzi*. Similarly, methionine can be converted to oxobutyrate, the precursor to propionyl-CoA, but no further. This could explain why threonine is not oxidized via the 2-oxobutyrate pathway in *T. brucei*: because the necessary threonine hydratase is missing. Instead, threonine is degraded to acetyl-CoA and glycine via the aminoacetone pathway by using a mitochondrial threonine dehydrogenase (*44*) and an aminoacetone synthase. The latter two genes are present in *T. brucei* and *T. cruzi* but absent from *L. major*.

Histidine catabolism appears to be absent from *T. brucei*. However, *T. cruzi* does have a pathway that looks typically eukaryotic, except for the last enzyme (glutamate formimino aminotransferase, Tc00.1047053507963.20), which only appears to be present in *T. cruzi* and *Tetrahymena*. Histidine is the precursor for ovothiol A, an antioxidant (*45*), which is in *L. major* and may help to protect the invading parasite from macrophage-derived hydrogen peroxide and nitric oxide. The biosynthetic pathway is thought to involve two intermediate steps requiring cysteine and adenosylmethionine. However, the sequences of these enzymes are not known in any organism and therefore cannot be identified.

A functional urea cycle (Plate 2) is missing from all three organisms. Although carbamoylphosphate synthetase is present in all three, this enzyme is also essential for pyrimidine biosynthesis. Argininosuccinate synthase and a bona fide glycosomal arginase were only identified in *L. major*. A second member of the arginase/agmatinase/formiminoglutamase gene family was found in all three parasites. However, deletion of the former gene in *L. major* renders it auxotrophic for ornithine, indicating that this second gene is probably not an arginase (*46*). Arginine kinase, present only in *T. brucei* and *T. cruzi*, has been proposed as a possible chemotherapeutic target (*47*). Phosphoarginine may play a role as a transient source of energy for renewal of ATP, similar to phosphocreatine in vertebrates.

**Trypanothione metabolism.** Arginine and ornithine are precursors for polyamine biosynthesis, essential for cell growth and differentiation and for the synthesis of the drug target, trypanothione (*48*). The first step in polyamine biosynthesis (ornithine decarboxylase) is the target for the human sleeping sickness drug difluoromethylornithine (Plate 2 boxed) (*48*) and was found in *T. brucei* and *L. major* but not *T. cruzi*. Arginine decarboxylase, catalyzing the proposed target for difluoromethylarginine in *T. cruzi* and an alternative route for putrescine synthesis, could not be identified either. Two candidate aminopropyltransferases were found in *T. cruzi*, in contrast to one in *L. major* and in *T. brucei*. The mammalian retroconversion pathway of polyamines appears to be absent.

Cysteine can be produced from homocysteine by the trans-sulfuration pathway present in all three organisms. De novo synthesis from serine appears possible only in *L. major* and *T. cruzi*. These parasites can also interconvert glycine and serine via serine hydroxymethyl transferase, which is apparently absent from *T. brucei*. Cysteine, glutamate, glycine, and methionine (as decarboxylated AdoMet) are the precursor amino acids for glutathione (GSH).

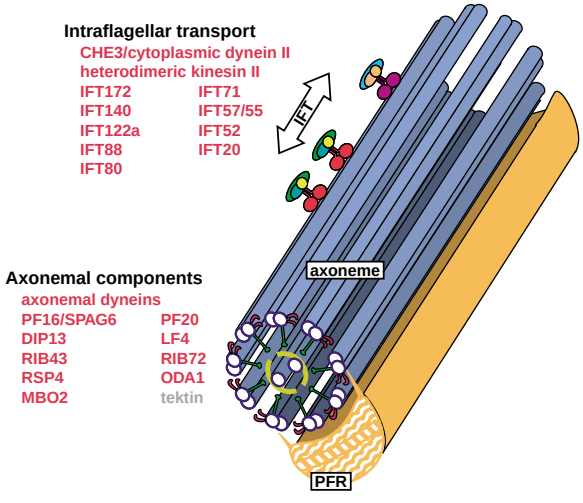


**Fig. 4.** Conservation of proteins of the flagellum in *T. brucei*, *T. cruzi*, and *L. major*. The figure shows the kinetoplastid flagellum with the axoneme (blue) and the lattice-like paraflagellar rod (PFR), unique to Kinetoplastida and Euglenida (yellow). Flagellar motors are seen attached to the axoneme. Components for which the Tritryps encode one or more homologs are indicated in red text. Those that do not appear to be encoded by the three genomes are indicated in gray text. (IFT, intraflagellar transport).

Trypanothione is synthesized from spermidine and glutathione and subsumes many of the roles of the latter in defending against chemical and oxidant stress. *T. cruzi* can also synthesize or salvage spermine to form the spermine-containing analog of trypanothione. Trivalent arsenical and antimonial drugs form conjugates with trypanothione and/or glutathione (*48*, *49*), and the synthesis of GSH, polyamines, and trypanothione are essential for survival in *T. brucei* (*50*–*52*).

The reactive by-product of metabolism, methylglyoxal, can be converted via the trypanothione-dependent glyoxalase pathway (*53*) to D-lactate. Homologs of glyoxalase I and II were found in *T. cruzi*, whereas *T. brucei* apparently lacks glyoxalase I but does have glyoxalase II (*54*). Unlike many bacteria, none of the parasites contain methylglyoxal synthase.

Trypanosomatids lack catalase and selenium-dependent peroxidases and are unusually dependent on trypanothione-dependent peroxidases for removal of peroxides. Tryparedoxins are homologs of thioredoxins, and all species contain both redox proteins. Thioredoxin reductase is absent, and all Tritryp peroxidases are coupled to trypanothione and trypanothione reductase, a validated drug target.

**Purine salvage and pyrimidine synthesis.** Several studies have shown that the trypanosomatids are incapable of de novo purine synthesis. This is now confirmed by the absence of 9 of the 10 genes required to make inosine monophosphate (IMP) from phosphoribosyl pyrophosphate (PRPP). Adenylosuccinate lyase is present, but this plays a role in purine salvage converting IMP to AMP. Present also are a large number of nucleobase and nucleoside transporters (*55*), (table S9) numerous enzymes for the interconversion of purine bases, and nucleosides and enzymes to synthesize pyrimidines de novo.

**Lipid and sterol metabolism.** Sterol metabolism has attracted considerable interest as a drug target in *T. cruzi* and *L. major* (*56*). Except for bloodstream *T. brucei*, which obtains cholesterol from the host, these parasites synthesize ergosterol and are susceptible to antifungal agents that inhibit this pathway. Candidate genes for most of ergosterol biosynthesis are present in all three parasites. Intermediates of this pathway, namely farnesyl- and geranyl-pyrophoshosphate, are also precursors for dolichols and the side chain of ubiquinone. *L. major* is capable of synthesizing the aromatic ring of ubiquinone from acetate, parahydroxybenzoate being an important intermediate. In this behavior it resembles prokaryotes.

The Tritryps seem to be capable of oxidizing fatty acids via β-oxidation in two separate celular compartments: glycosomes and mitochondria. This contrasts with yeasts and plants, where β-oxidation of fatty acids takes place exclusively in peroxisomes. In addition, the three trypanosomatids are capable of fatty acid biosynthesis.

**Summary and concluding remarks.** After more than a century of research aimed at understanding, controlling, and treating African trypanosomiasis, the genome sequence represents a major step in this process. It will not provide immediate relief to those suffering from trypanosomiasis. Instead, it will accelerate trypanosomiasis research throughout the scientific community. Through early data release before completion of the sequencing, the Tritryp genome projects have already allowed scientists to identify many new drug targets and pathways (e.g., GPI biosynthesis and isoprenoid metabolism) and have provided a framework for functional studies. Because about 50% of the genes of all three parasites have no known function, many more biochemical pathways and structural functions await discovery.

### References and Notes

1. *The World Health Report 2004: Changing History* (World Health Organisation, Geneva, 2004).
2. B. Wickstead, K. Ersfeld, K. Gull, *Genome Res.* **14**, 1014 (2004).
3. S. E. Melville, V. Leech, C. S. Gerrard, A. Tait, J. M. Blackwell, *Mol. Biochem. Parasitol.* **94**, 155 (1998).
4. F. Bringaud *et al.*, *Eukaryot. Cell* **1**, 137 (2002).
5. Materials and methods are available as supporting material on *Science* Online.
6. N. Hall *et al.*, *Nucleic Acids Res.* **31**, 4864 (2003).
7. N. M. El-Sayed *et al.*, *Nucleic Acids Res.* **31**, 4856 (2003).
8. A. C. Ivens *et al.*, *Science* **309**, 436 (2005).
9. N. M. El-Sayed *et al.*, *Science* **309**, 404 (2005).
10. P. Borst, *Cell* **109**, 5 (2002).
11. E. Pays, L. Vanhamme, D. Perez-Morga, *Curr. Opin. Microbiol.* **7**, 369 (2004).
12. L. H. Van der Ploeg *et al.*, *Nucleic Acids Res.* **10**, 5905 (1982).
13. C. M. Turner, J. D. Barry, *Parasitology* **99**, 67 (1989).
14. A. Y. Liu, L. H. Van der Ploeg, F. A. Rijsewijk, P. Borst, *J. Mol. Biol.* **167**, 57 (1983).
15. A. Craig, A. Scherf, Eds., *Antigenic Variation* (Academic Press, Amsterdam, 2003), pp. 464.
16. C. Roth, F. Bringaud, R. E. Layden, T. Baltz, H. Eisen, *Proc. Natl. Acad. Sci. U.S.A.* **86**, 9375 (1989).
17. S. M. Kamper, A. F. Barbet, *Mol. Biochem. Parasitol.* **53**, 33 (1992).
18. I. R. Arkhipova, H. G. Morrison, *Proc. Natl. Acad. Sci. U.S.A.* **98**, 14497 (2001).
19. J. D. Barry, M. L. Ginger, P. Burton, R. McCulloch, *Int. J. Parasitol.* **33**, 29 (2003).
20. H. H. Kazazian Jr., *Science* **303**, 1626 (2004).
21. M. Carrington *et al.*, *J. Mol. Biol.* **221**, 823 (1991).
22. L. Vanhamme *et al.*, *Nature* **422**, 83 (2003).
23. M. J. Gardner *et al.*, *Nature* **419**, 498 (2002).
24. M. Becker *et al.*, *Genome Res.* **14**, 2319 (2004).
25. P. Overath, M. Engstler, *Mol. Microbiol.* **53**, 735 (2004).
26. M. C. Field, M. Carrington, *Traffic* **5**, 905 (2004).
27. J. P. Ackers, V. Dhir, M. C. Field, *Mol. Biochem. Parasitol.* **141**, 89 (2005).
28. H. P. Price, C. Panethymitaki, D. Goulding, D. F. Smith, *J. Cell Sci.* **118**, 831 (2005).
29. E. Ogbadoyi, K. Ersfeld, D. Robinson, T. Sherwin, K. Gull, *Chromosoma* **108**, 501 (2000).
30. J. E. Lowell, G. A. Cross, *J. Cell Sci.* **117**, 5937 (2004).
31. J. A. Garcia-Salcedo *et al.*, *EMBO J.* **23**, 780 (2004).
32. J. A. Atwood III *et al.*, *Science* **309**, 473 (2005).
33. J. J. Cazzulo, *FASEB J.* **6**, 3153 (1992).
34. S. Besteiro *et al.*, *J. Biol. Chem.* **277**, 38001 (2002).
35. J. J. Van Hellemond, F. R. Opperdoes, A. G. Tielens, *Proc. Natl. Acad. Sci. U.S.A.* **95**, 3036 (1998).
36. C. Nihei, Y. Fukai, K. Kita, *Biochim. Biophys. Acta* **1587**, 234 (2002).
37. M. A. Ferguson, *Proc. Natl. Acad. Sci. U.S.A.* **97**, 10673 (2000).
38. J. R. Roper, M. L. Guther, K. G. Milne, M. A. Ferguson, *Proc. Natl. Acad. Sci. U.S.A.* **99**, 5884 (2002).
39. M. E. Drew *et al.*, *Mol. Cell. Biol.* **15**, 5508 (1995).
40. R. F. Steiger, E. Steiger, *J. Protozool.* **24**, 437 (1977).
41. A. El Sawalhy, J. R. Seed, J. E. Hall, H. El Attar, *J. Parasitol.* **84**, 469 (1998).
42. V. Gazit, R. Ben-Abraham, C. G. Pick, Y. Katz, *Behav. Brain Res.* **143**, 1 (2003).
43. M. L. Ginger, M. L. Chance, I. H. Sadler, L. J. Goad, *J. Biol. Chem.* **276**, 11674 (2001).
44. D. J. Linstead, R. A. Klein, G. A. Cross, *J. Gen. Microbiol.* **101**, 243 (1977).
45. D. J. Steenkamp, *Antioxid. Redox Signal.* **4**, 105 (2002).
46. S. C. Roberts *et al.*, *J. Biol. Chem.* **279**, 23668 (2004).
47. C. A. Pereira, G. D. Alonso, H. N. Torres, M. M. Flawia, *J. Eukaryot. Microbiol.* **49**, 82 (2002).
48. A. H. Fairlamb, *Trends Parasitol.* **19**, 488 (2003).
49. S. Wyllie, M. L. Cunningham, A. H. Fairlamb, *J. Biol. Chem.* **279**, 39925 (2004).
50. T. T. Huynh, V. T. Huynh, M. A. Harmon, M. A. Phillips, *J. Biol. Chem.* **278**, 39794 (2003).
51. F. Li, S. B. Hua, C. C. Wang, K. M. Gottesdiener, *Exp. Parasitol.* **88**, 255 (1998).
52. M. A. Comini *et al.*, *Free Radic. Biol. Med.* **36**, 1289 (2004).
53. T. J. Vickers, N. Greig, A. H. Fairlamb, *Proc. Natl. Acad. Sci. U.S.A.* **101**, 13186 (2004).
54. T. Irsch, R. L. Krauth-Siegel, *J. Biol. Chem.* **279**, 22209 (2004).
55. S. M. Landfear, B. Ullman, N. S. Carter, M. A. Sanchez, *Eukaryot. Cell* **3**, 245 (2004).
56. J. A. Urbina, R. Docampo, *Trends Parasitol.* **19**, 495 (2003).
57. We thank our colleagues at the Wellcome Trust Sanger Institute and the Institute for Genomic Research, colleagues in the *T. brucei* and Tritryp genome networks, and the Trypanosomatid research community for their support. We also thank K. Stuart, P. Myler, E. Worthey [Seattle Biomedical Research Institute (SBRI)], and B. Andersson (Karolinska Institutet) for helpful discussions of the manuscript and P. Myler and G. Aggarwal (SBRI) for preparation of Plate 3. Funding for this work was provided by the Wellcome Trust and a grant from the National Institute for Allergy and Infectious Diseases (NIAID) to N.M.E.-S. (AI43062). Funding for provision of DNA resources was provided by Wellcome Trust grants to S.E.M. (062513) and C.M.R.T. (062513). Funding for several *T. brucei* genome network and Tritryp genome meetings over the past decade was provided by the Burroughs Wellcome Fund, NIAID, the Wellcome Trust, and the World Health Organisation Special Programme for Research and Training in Tropical Diseases. Correspondence and requests for sequence data should be addressed to M.B. (mb4@sanger.ac.uk) or N.M.E.-S. (nelsayed@tigr.org). Requests for DNA resources should be addressed to S.E.M. (sm160@cam.ac.uk), and requests for updates or amendments to genome annotation should be directed to C.H.-F. (chf@sanger.ac.uk). All data sets and the genome sequence and annotation are available through GeneDB at www.genedb.org. Sequence data have been deposited at DNA Data Bank of Japan/European Molecular Biology Laboratory/GenBank with consecutive accession numbers CP000066 to CP000071 for chromosomes 3 to 8 and project accession numbers AAGZ00000000, AAHA00000000, and AAHB0000000 for the whole-chromosome shotgun projects of chromosomes 9 to 11. The versions of chromosomes 9 to 11 described in this paper are the first versions, AAGZ01000000, AAHA01000000, and AAHB01000000, and unassembled contigs (overlapping contiguous sequences) have accession number CR940345.

**Supporting Online Material**

www.sciencemag.org/cgi/content/full/309/5733/416/DC1
Materials and Methods
Figs. S1 to S5
Tables S1 to S11
References and Notes

23 March 2005; accepted 22 June 2005
10.1126/science.1112642