The Genome Sequence of *Trypanosoma cruzi*, Etiologic Agent of Chagas Disease

Najib M. El-Sayed,[1,2*†] Peter J. Myler,[3,4,5*†] Daniella C. Bartholomeu,[1] Daniel Nilsson,[6] Gautam Aggarwal,[3] Anh-Nhi Tran,[6] Elodie Ghedin,[1,2] Elizabeth A. Worthey,[3] Arthur L. Delcher,[1] Gaëlle Blandin,[1] Scott J. Westenberger,[1,7] Elisabet Caler,[1] Gustavo C. Cerqueira,[1,8] Carole Branche,[6] Brian Haas,[1] Atashi Anupama,[3] Erik Arner,[6] Lena Åslund,[9] Philip Attipoe,[3] Esteban Bontempi,[6,10] Frédéric Bringaud,[11] Peter Burton,[12] Eithon Cadag,[3] David A. Campbell,[7] Mark Carrington,[13] Jonathan Crabtree,[1] Hamid Darban,[6] Jose Franco da Silveira,[14] Pieter de Jong,[15] Kimberly Edwards,[6] Paul T. Englund,[16] Gholam Fazelina,[3] Tamara Feldblyum,[1] Marcela Ferella,[6] Alberto Carlos Frasch,[17] Keith Gull,[18] David Horn,[19] Lihua Hou,[1] Yiting Huang,[3] Ellen Kindlund,[6] Michele Klingbeil,[20] Sindy Kluge,[6] Hean Koo,[1] Daniela Lacerda,[1,21] Mariano J. Levin,[22] Hernan Lorenzi,[22] Tin Louie,[3] Carlos Renato Machado,[8] Richard McCulloch,[12] Alan McKenna,[6] Yumi Mizuno,[6] Jeremy C. Mottram,[12] Siri Nelson,[3] Stephen Ochaya,[6] Kazutoyo Osoegawa,[15] Grace Pai,[1] Marilyn Parsons,[3,4] Martin Pentony,[3] Ulf Pettersson,[9] Mihai Pop,[1] Jose Luis Ramirez,[23] Joel Rinta,[3] Laura Robertson,[3] Steven L. Salzberg,[1] Daniel O. Sanchez,[17] Amber Seyler,[3] Reuben Sharma,[13] Jyoti Shetty,[1] Anjana J. Simpson,[1] Ellen Sisk,[3] Martti T. Tammi,[6,24] Rick Tarleton,[25] Santuza Teixeira,[8] Susan Van Aken,[1] Christy Vogt,[3] Pauline N. Ward,[12] Bill Wickstead,[18] Jennifer Wortman,[1] Owen White,[1] Claire M. Fraser,[1] Kenneth D. Stuart,[3,4] Björn Andersson[6†]

Whole-genome sequencing of the protozoan pathogen *Trypanosoma cruzi* revealed that the diploid genome contains a predicted 22,570 proteins encoded by genes, of which 12,570 represent allelic pairs. Over 50% of the genome consists of repeated sequences, such as retrotransposons and genes for large families of surface molecules, which include trans-sialidases, mucins, gp63s, and a large novel family (>1300 copies) of mucin-associated surface protein (MASP) genes. Analyses of the *T. cruzi*, *T. brucei*, and *Leishmania major* (Tritryp) genomes imply differences from other eukaryotes in DNA repair and initiation of replication and reflect their unusual mitochondrial DNA. Although the Tritryp lack several classes of signaling molecules, their kinomes contain a large and diverse set of protein kinases and phosphatases; their size and diversity imply previously unknown interactions and regulatory processes, which may be targets for intervention.

Trypanosoma cruzi causes Chagas disease in humans. Acute infection can be lethal, but the disease usually evolves into a chronic stage, accompanied in 25 to 30% of cases by severe debilitation and ultimately death. It is estimated that 16 to 18 million people are infected, primarily in Central and South America, with 21,000 deaths reported each year (*1*). *T. cruzi* is normally transmitted by reduviid bugs via the vector feces after a bug bite and also after blood transfusion. Attempts to develop vaccines for parasitic diseases have been futile, and there is a critical lack of methods for diagnosis and treatment.

The taxon *T. cruzi* contains two defined groups, *T. cruzi* I and *T. cruzi* II, as well as additional groups yet to receive a designation (*2*). *T. cruzi* I is associated with the silvatic transmission cycle and infection of marsupials (*3*). *T. cruzi* II consists of five related subgroups, termed IIa, IIb, IIc, IId, and IIe (*4*), and is associated with the domestic transmission cycle and infection of placental mammals

(*5*). *T. cruzi* strain CL Brener is a member of subgroup IIe and was chosen for genome sequencing because it is well characterized experimentally (*6*). *T. cruzi* is heterozygous at many loci (*7*), with different-sized homologous chromosome pairs (*8*). Data from several laboratories (*9–13*) are consistent with its being a hybrid between subgroup IIb and subgroup IIc [which itself is also apparently a hybrid derived from *T. cruzi* I (*12*)]. The finding of *T. cruzi* I sequences in the CL Brener strain (*14*) further supports the role of multiple progenitors in the evolution of *T. cruzi* hybrid strains.

[1]Department of Parasite Genomics, The Institute for Genomic Research, Rockville, MD 20850, USA. [2]Department of Microbiology and Tropical Medicine, George Washington University, Washington, DC 20052, USA. [3]Seattle Biomedical Research Institute, Seattle, WA 98109, USA. [4]Department of Pathobiology, School of Public Health and Community Medicine, [5]Department of Medical Education and Biomedical Informatics, University of Washington, Seattle, WA 98195, USA. [6]Center for Genomics and Bioinformatics, Karolinska Institutet, Berzelius väg 35, S-171 77 Stockholm, Sweden. [7]Department of Microbiology, Immunology and Molecular Genetics, University of California, Los Angeles, CA 90095, USA. [8]Departamento de Bioquímica e Imunologia, Universidade Federal de Minas Gerais, CEP 31270-901, Belo Horizonte, MG, Brazil. [9]Department of Genetics and Pathology, Uppsala University, SE-751 85 Uppsala, Sweden. [10]Instituto Nacional de Parasitología Dr. M. Fatala Chabén, Administración Nacional de Laboratories e Insitutos de Salud (ANLIS), 1063, Buenos Aires, Argentina. [11]Laboratoire de Génomique Fonctionnelle des Trypanosomatides, UMR-CNRS 5162, Université Victor Segalen Bordeaux II, 33076 Bordeaux Cedex, France. [12]Wellcome Centre for Molecular Parasitology, University of Glasgow, Glasgow G11 6NU, Scotland, UK. [13]Department of Biochemistry, University of Cambridge, Cambridge CB2 1GA, UK. [14]Departamento de Microbiologia, Imunologia e Parasitologia, Universidade Federal de São Paulo, CEP 04023–062, São Paulo, SP, Brazil. [15]BACPAC Resources, Children's Hospital Oakland Research Institute, Oakland, CA 94609, USA. [16]Department of Biological Chemistry, Johns Hopkins University School of Medicine, Baltimore, MD 21205, USA. [17]Instituto de Investigaciones Biotecnológicas–Instituto Tecnológico de Chascómús, National University of San Martin and National Research Council, 1650 Buenos Aires, Argentina. [18]Sir William Dunn School of Pathology, University of Oxford, South Parks Road, Oxford, OX1 3RE, UK. [19]London School of Hygiene and Tropical Medicine, Keppel Street, London, WC1E 7HT, UK. [20]Department of Microbiology, University of Massachusetts, Amherst, MA 01003, USA. [21]René Rachou Research Center/CPqRR, Oswaldo Cruz Foundation, Belo Horizonte, MG, Brazil. [22]Laboratory of Molecular Biology of Chagas Disease, Instituto de Investigaciones en Ingeniería Genética y Biología Molecular, National Research Council (CONICET-CYTED project), School of Sciences, Centro de Genomica Aplicada-CeGA-University of Buenos Aires, 1428 Buenos Aires, Argentina. [23]Instituto de Biología Experimental, Universidad Central de Venezuela and ADEA-MCT, 1041-A Caracas, Venezuela. [24]Departments of Biological Sciences and Biochemistry, National University of Singapore, Singapore. [25]Center for Tropical and Emerging Global Diseases, Department of Cellular Biology, University of Georgia, Athens, GA 30602, USA.

*These authors contributed equally to this work.
†To whom correspondence should be addressed. E-mail: nelsayed@tigr.org (N.M.E.-S.); peter.myler@sbri.org (P.J.M.); bjorn.andersson@cgb.ki.se (B.A.)

In this research article, we report on the sequencing of the *T. cruzi* genome, with an emphasis on our analysis of the Tritryp kinome, DNA replication and repair machinery, and organization of retroelements, as well as surface proteins, in *T. cruzi*. Other aspects of trypanosomatid biology and new insights gained from sequencing the Tritryp genomes are discussed in the accompanying papers (*15–17*).

Genome sequencing, assembly, and annotation. The sequence was obtained by using the whole-genome shotgun (WGS) technique (table S1), because the high repeat content (>50%) and hybrid nature of the genome limited the initial "map-as-you-go" bacterial artificial chromosome (BAC) clone–based approach. Assembly parameters were modified to contend with the high allelic variation, and postassembly generation of 2.5× genome sequence coverage of the Esmeraldo strain from the progenitor subgroup IIb allowed us to distinguish the two haplotypes (*18*).

The current *T. cruzi* genome assembly consists of 5489 scaffolds (containing 8740 contigs) totaling 67 Mb. On the basis of the assembly results, the *T. cruzi* diploid genome size was estimated to be between 106.4 and 110.7 Mb, which is larger than the previous estimate of 87 Mb, based on densitometric analysis of pulse-field gel-separated chromosomal DNA (*19*). Analysis of the 60.4-Mb annotated dataset (Table 1 and table S3) revealed that 30.5 Mb contain sequence found at least twice in the assembly, which suggests that they likely represent the two different haplotypes in the *T. cruzi* CL Brener genome. Comparison of the contigs with reads from the Esmeraldo genome, which is a member of one of the progenitor subgroups (IIb), allowed us to distinguish the two haplotypes (*18*). The two haplotypes display high levels of gene synteny, with most differences because of insertion/deletions in intergenic and subtelomeric regions and/or amplification of repetitive sequences (Plate 1). The average sequence divergence between the two haplotypes is 5.4%, and the protein-coding regions are considerably more conserved (2.2% difference) than intergenic regions.

On the basis of our haplotype analyses, we estimate that the haploid *T. cruzi* genome contains about 12,000 genes (see table S2 for details). Automated analysis of the 4008 *T. cruzi* contigs using AUTOMAGI (*18*) initially predicted 25,013 protein-coding genes in the diploid genome, which was manually refined to a total of 22,570 genes, of which 6159 represent alleles present in the IIb haplotype, 6043 represent alleles from the other haplotype, and 10,368 represent sequences that could not be assigned to a particular haplotype (table S2). A total of 594 RNA genes were also identified from this same sequence dataset (Table 1 and table S3), although another 1400 were identified in the unannotated contigs, which contained many tandemly repeated ribosomal RNA (rRNA), spliced leader (sl) RNA, and small nucleolar RNA (snoRNA) genes. As seen in the other trypanosomatids, the protein-coding genes are generally arranged in long clusters of tens-to-hundreds of genes on the same DNA strand. Putative function could be assigned to 50.8% of the predicted protein-coding genes on the basis of significant similarity to previously characterized proteins or known functional domains (table S3).

Repeats, retrotransposons, and telomeres. At least 50% of the *T. cruzi* genome is repetitive sequence, consisting mostly of large gene families of surface proteins, retrotransposons, and subtelomeric repeats. TRIBE-MCL analysis [which uses the Markov cluster (MCL) algorithm] (*18*) revealed 1052 paralogous clusters (of more than two genes) encompassing 8419 genes, of which 46 clusters (3836 genes) contained 20 or more paralogues (table S5). The largest gene families (which often fall into several TRIBE-MCL clusters) encode surface proteins such as mucin-associated surface proteins (MASPs), members of the trans-sialidase (TS) superfamily, mucins, and the surface glycoprotein gp63 protease (Table 2) that are often *T. cruzi*–specific and account for ~18% of the total of protein-coding genes.

Table 1. Summary of the *T. cruzi* annotated genome. For RNA genes, see details in table S3. tRNA, transfer RNA; snRNA, small nuclear RNA; srpRNA, signal recognition particle RNA.

Parameter	Number
The genome	
Size* (bp)	60,372,297
G+C content (%)	51
Sequence scaffolds†	838
Sequence contigs	4,008
Percent coding	58.9
Protein-coding genes	
No. of gene models	23,216
No. of genes‡	22,570
Estimated no. of genes per haploid genome§	~12,000
Pseudogenes	3,590
Mean CDS length‖ (bp)	1,513
Median CDS length‖ (bp)	1,152
G+C content (%)	53.4
Gene density (genes per Mb)	385
Intergenic regions¶	
Mean length (bp)	1,024
G+C content (%)	47
RNA genes	
tRNA	115
rRNA	219
slRNA	192
snRNA	19
snoRNA	1,447
srpRNA	2

*Includes all scaffolds and contigs >5 kb, from both haplotypes. †784 scaffolds + 54 contigs. ‡Genes split across contig boundaries were counted once. §See details in table S2. ‖Excluding partial genes and pseudogenes. ¶Regions between protein-coding CDSs.

These genes occur in dispersed clusters of tandem and interspersed repeats, often at subtelomeric locations (see below). There is also a large family of β-galactofuranosyltransferases, which likely reflects the extensive use of glycoconjugates on the parasite cell surface, similar to that seen in *L. major* (*17*), but in contrast to *T. brucei*, which has many fewer genes encoding enzymes in the glycosylation pathway. In addition, a relatively large number of mostly housekeeping genes occur in highly conserved tandem clusters throughout the genome. Because similar gene organization is seen in *L. major* and *T. brucei*, one possible function of these repeats may be to increase the expression level of these proteins. The copy number of these genes is likely underestimated because of the collapse of multiple tandem repeats into fewer copies during assembly, as evidenced by regions of locally high sequence coverage (table S4). Interestingly, the degree of sequence conservation between repeat copies is generally higher within the same haplotype than between haplotypes, which suggests that the expansions are recent, or that specific mechanisms are in place to conserve the gene copies.

One example of gene family expansion has occurred in the kinetoplastid myosin genes. Analysis of the Tritryp genomes reveals two classes of myosin: conventional MyoI proteins and a novel family of kinetoplastid myosins. *T. brucei* and *L. major* have a single member of both families, but *T. cruzi* has expanded the kinetoplastid myosin family to seven (haploid) members (at dispersed loci) with a considerable diversity of sequence (Fig. 1). Moreover, *T. cruzi* has retained the CapZ F-actin capping complex that is absent in both *T. brucei* and *L. major* [see (*16*)], which suggests a difference in myosin function between the trypanosomatid species. It is possible that this may be associated with the cytostome-cytopharynx complex, the major cytoskeletal feature (a funnel-shaped invagination in the plasma membrane that is the site of endocytosis for macromolecules such as low density lipoprotein) found only in the Stercorarian trypanosomes (including *T. cruzi*) (*20*).

Long terminal repeat (LTR) and non-LTR retroelements account for ~5% of the haploid *T. cruzi* genome and 2% of the haploid *T. brucei* genome. Several copies of the site-specific non-LTR retrotransposons CZAR (*21*) and SLACS (*22*) are present in the SL RNA loci of *T. cruzi* and *T. brucei*, but absent from *L. major*. Although the autonomous *T. brucei ingi* (*23*) and *T. cruzi* L1*Tc* (*24*) non-LTR retroelements have been reported as randomly distributed in the host genome, analysis of their genomic context in the complete *T. brucei* and *T. cruzi* genomes indicates that they are preferentially inserted downstream of conserved **A**xxxxxxx**TtgxGTxGGxT**xxx↑**tTxTxxT**xxxxxxx↑ and **GA**xx**AxGa**xxxxx**txTATG**↑**A**xxxxxxxxxxx↑ motifs, respectively, where the arrows indicate the single-strand cleavage sites (*25*). The nonautonomous RIME (*26*) and NARTc (*27*) elements, respectively, are preceded by the same conserved motifs, which indicates that they likely use the *ingi* and L1*Tc* retrotransposition machinery. It is interesting that both retroelement pairs share their 5′ extremities, and the *ingi*/RIME pair also share their 3′ sequences. To our knowledge, this sequence conservation has never been reported for other LINE-/SINE-like couples. A similar situation was seen with the *T. brucei* and *T. cruzi* LTR-retrotransposon pairs (autonomous VIPER and nonautonomous SIRE).

In contrast to *T. brucei* and *T. cruzi* (which contain three potentially active *ingi* and 15 L1*Tc* elements, respectively), no active retrotransposons have been described in *Leishmania* species, but the *L. major* genome does contain 52 copies of a degenerate retroelement called "DIRE" (*28*) (Table 3). Phylogenetic analysis conducted on the reverse transcriptase domain of non-LTR retrotransposons from different eukaryotes indicates that *ingi*, L1*Tc*, and DIRE form a monophyletic group, which suggests that the common ancestor of trypanosomatids contained active retrotransposons that evolved into the presently active elements in *T. brucei*. The last active *L. major* retroelements were probably lost in the ancient past, and only their vestiges (DIREs) still reside in the present genome. In nematodes, the RNA interference (RNAi) machinery down-regulates retroelement mobilization, which prevents the negative effects of rampant expansion (*29*). It is noteworthy that RNAi is operational in *T. brucei*, whereas *T. cruzi* and *L. major* do not seem to have the full RNAi machinery (*15*, *30*, *31*). This suggests that *T. cruzi*, and perhaps other trypanosomatids, uses an alternative strategy for retroelement silencing.

In several protozoan parasites, the subtelomeric regions are often associated with large gene families encoding surface proteins. Analysis of the *T. cruzi* genome assembly reported here reveals 49 scaffolds that contain the terminal THR sequence, which indicates that they are likely telomeric (table S6). In most cases, the THR sequences are immediately adjacent to a 0.4- to 1.8-kb telomere-associated sequence (TAS), similar to the *T. cruzi*–specific 189–base pair (bp) junction described previously (*32*). The subtelomeric region between TAS and the first upstream nonrepetitive gene is characterized by a polymorphic assembly of RHS (retrotransposon hotspot) (*33*), TS superfamily (*34*), DGF-1 (dispersed gene family–1) (*35*) genes or pseudogenes, as well as VIPER/SIRE, L1*Tc*/NAR*Tc*, and/or DIRE retroelements. These genes are all on the same strand, such that they would be transcribed toward the telomere.

Telomerase activity has been reported in the Tritryps (*36*), and we have now identified the gene encoding the protein component

Table 2. Large gene families in *T. cruzi*. Members are listed as total genes (pseudogenes in parentheses).

Gene product	Members	Tritryp orthologs
trans-Sialidase (TS)	1430 (693)	*Tb*
MASP	1377 (433)	No
Mucin	863 (201)	No
Retrotransposon hot spot (RHS) protein	752 (557)	*Tb*
Dispersed gene family protein 1 (DGF-1)	565 (136)	No
Surface protease (gp63)	425 (251)	*Tb* + *Lm*
Mucinlike protein	123	No
Hypothetical	117	*Lm*+*Tb*
Hypothetical	93	*Lm*+*Tb*
Kinesin, putative	79	*Lm*+*Tb*
Protein kinase (CMGC group)	77	*Lm*+*Tb*
Protein kinase (several groups)	79	*Lm*+*Tb*
Hypothetical protein	42	No
Glycosyltransferase	52	*Lm*+*Tb*
RNA helicase (eIF-4a)	47	*Lm*+*Tb*
Protein kinase (NEK group)	39	*Lm*+*Tb*
MASP-related	38	No
Glycosyltransferase	36	*Lm*+*Tb*
Hypothetical	35	*Lm*+*Tb*
Amino acid permease	28	*Lm*+*Tb*
AAA ATPase	33	*Lm*+*Tb*
Protein phosphatase	30	*Lm*+*Tb*
Heat shock protein HSP70	21	*Lm*+*Tb*
Protein kinase (STE group)	25	*Lm*+*Tb*
RNA helicase	23	*Lm*+*Tb*
Phosphatidylinositol phosphate kinase–related	23	*Lm*+*Tb*
Hypothetical	24	*Lm*+*Tb*
Elongation factor 1-γ (EF-1-γ)	22	*Lm*+*Tb*
DNA helicase (DNA repair)	21	*Lm*+*Tb*
Actin-related	20	*Lm*+*Tb*
Cysteine peptidase	20	*Lm*+*Tb*

(TERT) of this enzyme in all three trypanosomatids (table S7), along with a putative homolog of telomerase-associated protein TEP1, which has been shown to interact with telomerase in other cell types (*37*). We were unable to identify other putative capping or telomere repeat–binding proteins, but all three trypanosomatids contain genes encoding two proteins (JBP1 and JBP2) that bind the β-D-glucosyl(hydroxymethyl)uracil DNA base modification (also known as J) enriched at telomeres in the bloodstream from *T. brucei* and in other trypanosomatids (*38*). JBP2 also has a snf2-like helicase domain, which suggests a possible role in gene regulation.

DNA repair, recombination, replication, and meiosis. The genes that encode many of the enzymatic components of DNA repair were identified in the *T. cruzi* (and Tritryp) genomes (table S8), and thus, these organisms appear able to catalyze most repair pathways. Three pathways of direct repair are apparent. Single homologs of *O*-6 methylguanine alkyltransferase, for alkylation reversal, and the AlkB dioxygenase, for oxidative damage repair, are present in all three genomes. However, *T. cruzi* does not contain a clear photolyase homolog, although *T. brucei* and *L. major* do, presumably for photoreactivation.

Most components of the base-excision repair pathway are conserved in the Tritryps. In contrast, genes implicated in mechanisms that prevent the effects of oxidative stress, such as catalase and the Mut T homolog 8-oxoguanine hydrolase, were not detected in any of the three parasites. About half of the DNA glycosylases described in other organisms are identifiable, but only one has been experimentally characterized (*39*). Trypanosomatids contain most components of the eukaryotic nucleotide excision repair pathway but may share some biochemical novelties with the less-characterized systems of plants and *Plasmodium falciparum*. For example, although the core XPB/RAD25 and XPD/RAD3 helicases, as well as the XPG/RAD2, XPF/RAD1, and ERCC1 endonucleases, are discernible, many other genes (including XPA/RAD14) are not. The possible consequences of these differences are unknown. All three genomes appear to contain a complete complement of genes for base mismatch repair.

Homologous recombination has been well documented in the trypanosomatids, because it is exploited for experimental genome manipulation (*40*) and is a key mechanism for antigenic variation that *T. brucei* uses for immune evasion (*41*). However, some genes for homologous recombination are notably absent, including RAD52, which is critical for homologous recombination in *Saccharomyces cerevisiae* (*42*). Surprisingly, the enzymatic machinery for nonhomologous end-joining is not readily detectable in the trypanosomatids, although homologs of KU70 and KU80, the components of the Ku heterodimer, were found and are known to function in *T. brucei* telomere length regulation (*43*). Thus, this enzymatic pathway may have been lost or altered in the trypanosomatids during evolution, as in *P. falciparum* (*44*). Multigene families encoding DNA polymerase κ were discovered in the three trypanosomatids. This enzyme is a low-fidelity, exonuclease-deficient DNA polymerase involved in translesion DNA synthesis.

The replication fork synthetic machinery of kinetoplastid nuclear chromosomes appears to resemble that in higher eukaryotes (table S9), although the machinery for initiation of replication may differ significantly. Most strikingly, Tritryps have a candidate gene for only one of the six subunits of the origin recognition complex, ORC1, which is also homologous to CDC6. Also, there are no clear orthologs for the MCM10, CDT1, DBF4, and possibly CDC7, proteins that play key roles in initiation of replication in *S. cerevisiae* and other eukaryotes (*45*). On the basis of the proteins encoded in the kinetoplastid genomes, replication initiation may resemble that in the Archaea, which also have only a single ORC subunit, ORC1/CDC6 (*46*), and lack the collection of initiation factors utilized by eukaryotes.

The Tritryp mitochondrial DNA is a unique network structure, known as kinetoplast DNA (kDNA), composed of thousands of minicircles and dozens of maxicircles topologically interlocked and replicated at a specific time in the cell cycle (*47*). The complexity of this structure dictates an unusual replication mechanism and accounts for the substantial differences from higher eukaryotes we observe. The Tritryp nuclear genomes encode six DNA polymerases that have been localized to the mitochondria in *T. brucei* (*48*, *49*), whereas yeast and mammalian mitochondria have only one, DNA polymerase γ. There also appear to be multiple DNA ligases (*50*) and helicases. The Tritryp genomes reveal no candidate genes for mitochondrial primase, single-strand binding protein, or DNA polymerase processivity factors, which suggests that these genes may have diverged from their prokaryotic or eukaryotic counterparts. In contrast, the gene for mitochondrial RNA polymerase, which apparently plays a role in maxicircle replication (*51*), resembles those from yeast and human. Finally, the Tritryp genomes provide no clues to the mechanisms triggering the initiation of kinetoplast DNA replication in a cell cycle–dependent manner.

Table 3. Retrotransposon copy numbers in the Tritryp genomes. The copy number per haploid genome is indicated, with the number of intact copies in parentheses.

Retrotransposons	Tb	Tc	Lm
LTR retrotransposons			
VIPER (4.5 kb)	26 (0)	275 (0)	0
SIRE (0.43 kb)	10 (0)	480 (0)	0
Non-LTR retrotransposons			
SLACS (6.3 kb)	4 (1)	0	0
CZAR (7.25 kb)	0	8 (*)	0
ingi (5.2 kb)	115 (3)	0	0
RIME (0.5 kb)	86 (0)	0	0
L1*Tc* (4.9 kb)	0	320 (15)	0
NAR*Tc* (0.25 kb)	0	133 (0)	0
DIRE(4 to 5 kb)	73 (0)	257 (0)	52 (0)

*The number of intact copies was not determined.

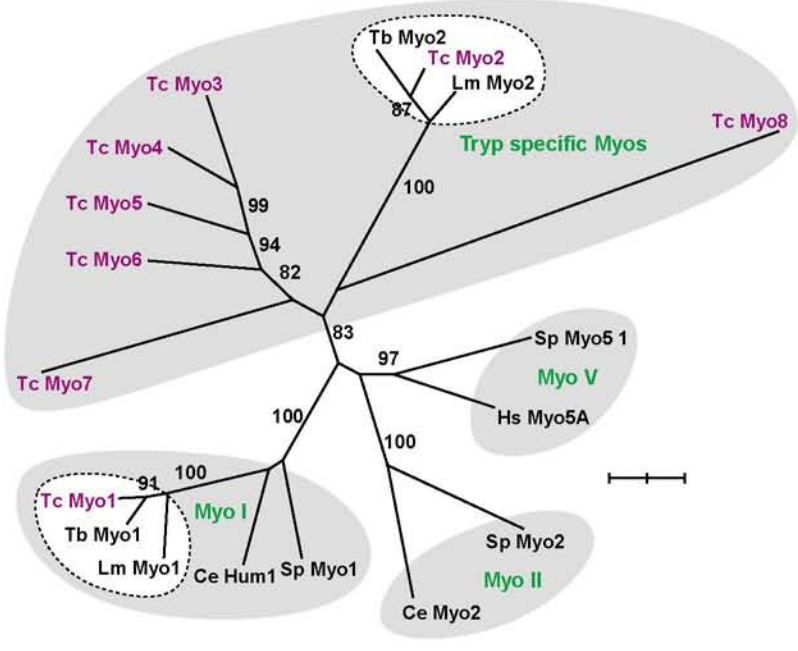

Fig. 1. Evolutionary analysis of trypanosomatid myosins, in comparison with myosins from *Schizosaccharomyces pombe* (Sp), *C. elegans* (Ce), and *Homo sapiens* (Hs). See supporting online material (*18*) for details.

The Tritryps are essentially diploid, but sexual reproduction is not an obligatory part of their life cycles. Genetic exchange occurs in both *T. brucei* (*52*) and *T. cruzi* (*11*), and there is Mendelian inheritance in *T. brucei* (*53*), but the molecular processes involved in genetic exchange are poorly characterized. Of the genes uniquely expressed during meiosis, six were identified as being sufficiently conserved to be readily identifiable in the genomes of most eukaryotes with an obligatory meiosis. Homologs for SPO11, DMC1, MND1, MSH4 (except for *L. major*), and MSH5, which are involved in homologous recombination, and HOP1, which is a component of the lateral elements of the synaptonemal complex, were identified in the Tritryp genomes. Thus, the Tritryps have the potential to undergo meiotic homologous exchange, but it is not possible to determine whether the potential for reductive divisions is present.

Signaling pathways. Several classes of important signaling molecules are absent in trypanosomatids, including serpentine receptors, heterotrimeric G proteins, most classes of catalytic receptors, SH2 and SH3 interaction domains, and regulatory transcription factors. Some catalytic receptors have been found, and all are adenylate cyclases (47 genes in *T. brucei*, 11 in *L. major*, and 25 in *T. cruzi*). The Tritryps, however, have a large and complex set of protein kinases (PKs), as well as a diversity of protein phosphatases (tables S10 and S11). They also have multiple enzymes involved in phosphoinositide metabolism, as well as modular domains that interact with those small molecules, although little is known concerning their functions (table S12).

Table 4. Comparison of Tritryp, yeast, and human kinome. The comparison is based on catalytic domains. Data for human (Hs) and yeast (Sc) were derived from Manning *et al.* (*68*).

PKs	Tb	Tc	Lm	Sc	Hs
		Eukaryotic PKs			
AGC	12	12	11	17	63
CAMK	13	13	16	21	74
CK1	5	6	7	4	12
CMGC	42	41*	47	21	61
NEK	20	23	22	1	15
STE	24	28	32	14	47
TK	0	0	0	0	90
TKL	0	0	0	0	43
Unique	21	25	27	4	7
Other†	19	19	18	33	61
Total	156	167	180	115	478
		Atypical PKs			
ABC	5	5	5	3	5
Alpha	2	1	5	0	6
Bud32	1	1	1	1	1
Cofilin	1	1	1	1	2
PDHK	3	3	3	2	5
PIKK	6	6	6	5	6
RIO	2	2	2	2	3
Other	0	0	0	1	12
Total	20	19	23	15	40

*Multiple copies of CDK8 in Tc were counted as one gene. †Other trypanosomatid eukaryotic PKs include Aurora, CAMKK, CK2, PEK, PLK, TLK, ULK, VPS15, and WEE1 kinases.

The Tritryp genomes encode 180, 156, and 167 distinct eukaryotic PKs in *L. major*, *T. brucei*, and *T. cruzi*, respectively, that are likely to be catalytically active, as well as 23, 20, and 19 atypical PKs, respectively. The trypanosomatid kinome is more than twice that of *P. falciparum* (*54*) and one-third larger than that of *S. cerevisiae*, although the overall representation of PK groups is similar (Table 4). Almost all receptor PKs in mammals are tyrosine kinases, but no such group is present in the parasites, and only a handful of trypanosomatid PKs have predicted transmembrane domains. Indeed, catalytic domains that map to the tyrosine kinase group are entirely missing, but dual-specificity kinases are present. Also missing is the TKL group, which shows features of both tyrosine and serine-threonine kinases, and the RGC (receptor guanylate cyclase) group, which is structurally related to PKs. The expansion of a few groups of PKs hints at a regulatory complexity focused on stress and the cell cycle (*55*). For example, *T. brucei* has many CMGC PKs, including 11 cyclin-dependent kinases (CDKs) (plus 10 cyclins) and 11 mitogen-activated protein kinases (MAP kinases). The STE kinases, which function in the MAP kinase activation cascade, are also very numerous in trypanosomatids. Their roles in trypanosomatids are relatively unexplored, although the importance of signaling pathways in flagellar length control (*56*), differentiation (*57*), and cellular proliferation (*58*) is apparent. A further example of expansion is the large NEK family of trypanosomatids. More than 20 PKs could not be classified into any established group. Their low similarity to human PKs raises the possibility of targeted intervention.

A key finding is the relative lack of identifiable accessory domains on the trypanosomatid PKs. Although ~50% of human PKs bear an additional PFAM domain, only 14% of Tritryp predicted PKs do. The diversity of such domains is also more restricted, with 83 domains represented on human PKs but only 21 represented on Tritryp PKs. Most striking in their absence are the domains most frequently found on human PKs: SH2, SH3, FN-III, and immunoglobulin-like domains. Most well represented in trypanosomatid PKs are PH domains (with six or seven in each species), again pointing to a role of phosphoinositide metabolism in regulatory networks in the parasite. Despite the paucity of recognizable domains, the vast majority of Tritryp PKs are much larger than a simple catalytic domain.

Surface molecules. Many trypanosomatid surface proteins are heavily glycosylated. Although the Tritryps have biosynthetic pathways for some sugars and have several glycosyltransferases (*17*), they are unable to synthesize sialic acid, which is present in several parasite surface glycoconjugates. However, in *T. cruzi* and *T. brucei*, incorporation of host sialic acid is possible because of a surface-bound TS (*34*, *59*), which can transfer sialidase from sialoglycoconjugates in the host to the terminal β-galactose on the highly *O*-glycosylated mucins in *T. cruzi* (*34*), and the glycosylphosphatidylinositol (GPI) anchor of procyclin in the insect stage of *T. brucei* (*60*).

Intriguingly, in comparison with *L. major* and *T. brucei*, *T. cruzi* shows a dramatic expansion of several families of surface molecules, including the TS, mucin, MASP, and gp63 protease families, which are each encoded by several hundred genes in the *T. cruzi* genome (Table 2). The *T. cruzi* assembly contains 1430 gene members of the TS superfamily, including 693 pseudogenes, which have previously been classified into two major subfamilies (*34*) (table S13). One subfamily includes 12 genes that share more than 90% identity with genes encoding enzymatically active TSs. This number may represent an underestimate because of collapsed assembly of near-identical repeats. Most, but not all, active TSs contain a variable number of 12–amino acid SAPA (shed acute-phase antigen) repeats and are GPI-anchored (*61*, *62*). The remaining TS superfamily members consist of more than 725 genes encoding enzymatically inactive TS-like proteins with variable degrees of homology to the active TSs. Only 371 genes have the conserved sialidase superfamily motif (VTVxNVxLYNR). The significant sequence variability suggests a strong selective pressure on the TS gene family to diversify. This pressure may be in part provided by the mammalian immune response, because TSs are targets of both humoral and cell-mediated immune responses (*34*). The TS family is much smaller in *T. brucei* and is absent from *L. major*.

The mucins represent another large (863 members) family of surface molecules in *T. cruzi*, which can be divided into two subfamilies (table S14). The 19-member TcSMUG family is relatively homogeneous, and members are expressed in the epimastigote stage in the insect vector (*63*). The much larger TcMUC subfamily is expressed in the mammalian stages (*64*) and contains 844 members. No mucin-related genes are found in *T. brucei*, but eight members of the large PSA-2 family (*17*) in *L. major* have a structure (including T_7KP_2 repeats) similar to those of TcMUC group I.

As indicated above, the TS genes can be found in subtelomeric repetitive regions, although they also occur in intrachromosomal arrays, often at Tritryp synteny breaks [see (*15*)]. We have identified another large *T. cruzi*–specific gene family within large (up to 600 kb) clusters of TS and mucin genes,

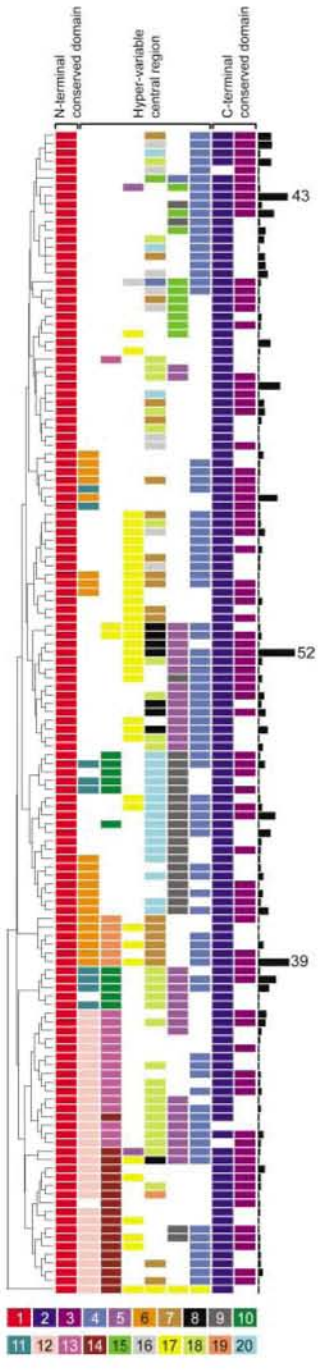

Fig. 2. Schematic representation of MASP protein structure and variability. The MEME algorithm (version 3.0) was used to identify motifs shared by members of the MASP family. The relative numbers in each of the defined subgroups are represented as a histogram on the right. The N- and C-terminal conserved domains and the central variable region are indicated on the top. Because patterns of variable length cause gaps and are split by MEME into two or more separate motifs, the C-terminal conserved domain of MASP is represented by two motifs separated by variably repeated leucine and valine residues. The motif consensus sequences are numbered in decreasing order of statistical significance and color coded. The MEME parameters, grouping methods, and amino acid sequence corresponding to each of the motifs are listed in the supporting online material (*18*).

members of which are characterized by conserved N- and C-terminal domains that encode a signal peptide and a GPI anchor addition site, respectively, which suggests a surface location in the parasite. The central region of these proteins is highly variable (Fig. 2) and often contains repeated sequence. Because most members of this family are located downstream of TcMUC II mucins (which they resemble structurally, if not at the sequence level), we have named the family mucin-associated surface proteins (MASPs). Of the 1377 *masp* genes identified, 771 appear to be intact and encode both N- and C-terminal conserved regions; 433 are pseudogenes. An interesting observation is the existence of chimeras (*26*) that contain the N- or C-terminal conserved domain of MASP combined with the N- or C-terminal domain of mucin or the C-terminal domain from the TS superfamily. The mechanism for the generation of such chimeric *masp* genes is unknown, although previous studies have described mosaic genes formed by group II and III members of the TS superfamily (*65*). Proteomic data from four different *T. cruzi* developmental stages revealed at least four distinct *masp* genes in trypomastigotes and another in epimastigotes (*66*). The low number of MASP peptides detected by proteomic approaches suggests that MASPs may contain extensive posttranslational modifications. Alternatively, *masp* genes may be expressed in intermediate stages not represented in the proteome data or may be expressed in a mutually exclusive fashion, similar to the *T. brucei* variant surface glycoproteins (VSGs).

The gp63 family of surface metalloproteases is found in the three trypanosomatids and has been implicated in virulence, host cell infection, and release of parasite surface proteins (*67*). Although *L. major* has only four gp63 genes and two gp63-like genes, and *T. brucei* has only 13, *T. cruzi* contains more than 420 genes and pseudogenes. These appear to be dispersed throughout the genome, although they sometimes occur in tandem clusters. The reason for this massive expansion of the gp63 gene family in *T. cruzi* is not yet apparent.

Several common themes emerge from genomic examination of Tritryp surface proteins: Many are highly glycosylated, and the proteins are members of large families containing highly variable central domains. The genes in *T. cruzi* and *T. brucei* are often located in large haploid arrays. It is likely that they have evolved to evade the host immune response, and the presence of pseudogenes may contribute to the diversity of the sequence repertoire through recombination. Nevertheless, species-specific differences do occur, because *T. brucei* expresses only one VSG at a time and has evolved a sophisticated system to constantly change the expressed copy, whereas *T. cruzi* simultaneously expresses numerous copies of the TSs, mucins, and probably MASPs and gp63s.

Implications for novel therapies. The elucidation of critical pathways in DNA repair, DNA replication, and meiosis and the identification of numerous protein kinases and phosphatases afforded by analysis of the Tritryp genomes promise to provide novel drug targets. Differences from the typical eukaryotic machinery for nucleotide excision/repair, initiation of DNA replication, and the presence of additional bacteria-like DNA polymerases used in replication of the mitochondrial genome all provide potential points of attack against the parasites. In addition, the presence of several PKs with little similarity to those in other eukaryotes present new possibilities for targeted drug development. The surface TS activity, which is, in *T. cruzi* at least, essential for incorporation of host sialic acid into parasite glycoconjugates, is another target for chemotherapeutic intervention, and work is already well advanced in this area (*58*). The elucidation of the complete repertoire of active *T. cruzi* TSs should help in this endeavor.

References and Notes

1. WHO, *The World Health Report, 2002* (World Health Organization, Geneva, 2002).
2. Anonymous, *Mem. Inst. Oswaldo Cruz* **94**, 429 (1999).
3. C. G. Clark, O. J. Pung, *Mol. Biochem. Parasitol.* **66**, 175 (1994).
4. S. Brisse, C. Barnabé, M. Tibayrenc, *Int. J. Parasitol.* **30**, 35 (2000).
5. M. R. Briones, R. P. Souto, B. S. Stolf, B. Zingales, *Mol. Biochem. Parasitol.* **104**, 219 (1999).
6. B. Zingales *et al.*, *Acta Trop.* **68**, 159 (1997).
7. N. R. Sturm, N. S. Vargas, S. J. Westenberger, B. Zingales, D. A. Campbell, *Int. J. Parasitol.* **33**, 269 (2003).
8. A. Pedroso, E. Cupolillo, B. Zingales, *Mol. Biochem. Parasitol.* **129**, 79 (2003).
9. S. Brisse *et al.*, *Mol. Biochem. Parasitol.* **92**, 253 (1998).
10. C. A. Machado, F. J. Ayala, *Proc. Natl. Acad. Sci. U.S.A.* **98**, 7396 (2001).
11. M. W. Gaunt *et al.*, *Nature* **421**, 936 (2003).
12. S. J. Westenberger, C. Barnabé, D. A. Campbell, N. R. Sturm, *Genetics*, in press.
13. S. Brisse *et al.*, *Infect. Genet. Evol.* **2**, 173 (2003).
14. M. C. Elias *et al.*, *Mol. Biochem. Parasitol.* **140**, 221 (2005).
15. N. M. El-Sayed *et al.*, *Science* **309**, 404 (2005).
16. M. Berriman *et al.*, *Science* **309**, 416 (2005).
17. A. C. Ivens *et al.*, *Science* **309**, 436 (2005).
18. Materials and methods are available as supporting material on *Science* Online.
19. M. I. Cano *et al.*, *Mol. Biochem. Parasitol.* **71**, 273 (1995).
20. W. De Souza, *Curr. Pharm. Des.* **8**, 269 (2002).
21. M. S. Villanueva, S. P. Williams, C. B. Beard, F. F. Richards, S. Aksoy, *Mol. Cell. Biol.* **11**, 6139 (1991).
22. S. Aksoy, T. M. Lalor, J. Martin, L. H. Van der Ploeg, F. F. Richards, *EMBO J.* **6**, 3819 (1987).
23. B. E. Kimmel, O. K. ole-MoiYoi, J. R. Young, *Mol. Cell. Biol.* **7**, 1465 (1987).
24. M. Olivares *et al.*, *Electrophoresis* **21**, 2973 (2000).
25. F. Bringaud *et al.*, *Mol. Biol. Evol.* **21**, 520 (2004).
26. G. Hasan, M. J. Turner, J. S. Cordingley, *Cell* **37**, 333 (1984).
27. F. Bringaud *et al.*, *Mol. Biochem. Parasitol.* **124**, 73 (2002).
28. E. Ghedin *et al.*, *Mol. Biochem. Parasitol.* **134**, 183 (2004).

29. N. L. Vastenhouw *et al.*, *Curr. Biol.* **13**, 1311 (2003).
30. W. D. DaRocha, K. Otsu, S. M. Teixeira, J. E. Donelson, *Mol. Biochem. Parasitol.* **133**, 175 (2004).
31. K. A. Robinson, S. M. Beverley, *Mol. Biochem. Parasitol.* **128**, 217 (2003).
32. M. A. Chiurillo, I. Cano, J. F. Da Silveira, J. L. Ramirez, *Mol. Biochem. Parasitol.* **100**, 173 (1999).
33. F. Bringaud *et al.*, *Eukaryot. Cell* **1**, 137 (2002).
34. A. C. Frasch, *Parasitol. Today* **16**, 282 (2000).
35. P. Wincker, A. C. Murto-Dovales, S. Goldenberg, *Mol. Biochem. Parasitol.* **55**, 217 (1992).
36. M. I. Cano, J. M. Dungan, N. Agabian, E. H. Blackburn, *Proc. Natl. Acad. Sci. U.S.A.* **96**, 3616 (1999).
37. Y. Liu *et al.*, *Mol. Cell. Biol.* **20**, 8178 (2000).
38. M. Cross *et al.*, *Mol. Microbiol.* **46**, 37 (2002).
39. J. Pena-Diaz *et al.*, *J. Mol. Biol.* **342**, 787 (2004).
40. S. M. Beverley, *Nat. Rev. Genet.* **4**, 11 (2003).
41. E. Pays, L. Vanhamme, D. Perez-Morga, *Curr. Opin. Microbiol.* **7**, 369 (2004).
42. L. S. Symington, *Microbiol. Mol. Biol. Rev.* **66**, 630 (2002).
43. C. Conway *et al.*, *J. Biol. Chem.* **277**, 21269 (2002).
44. M. J. Gardner *et al.*, *Nature* **419**, 498 (2002).
45. S. P. Bell, A. Dutta, *Annu. Rev. Biochem.* **71**, 333 (2002).
46. L. M. Kelman, Z. Kelman, *Mol. Microbiol.* **48**, 605 (2003).
47. R. Woodward, K. Gull, *J. Cell Sci.* **95**, 49 (1990).
48. M. M. Klingbeil, S. A. Motyka, P. T. Englund, *Mol. Cell* **10**, 175 (2002).
49. T. T. Saxowsky, G. Choudhary, M. M. Klingbeil, P. T. Englund, *J. Biol. Chem.* **278**, 49095 (2003).
50. K. M. Sinha, J. C. Hines, N. Downey, D. S. Ray, *Proc. Natl. Acad. Sci. U.S.A.* **101**, 4361 (2004).
51. J. Grams *et al.*, *J. Biol. Chem.* **277**, 16952 (2002).
52. L. Jenni *et al.*, *Nature* **322**, 173 (1986).
53. C. M. Turner *et al.*, *Parasitology* **129**, 445 (2004).
54. P. Ward, L. Equinet, J. Packer, C. Doerig, *BMC Genomics* **5**, 79 (2004).
55. T. C. Hammarton, J. C. Mottram, C. Doerig, *Prog. Cell Cycle Res.* **5**, 91 (2003).
56. M. Wiese, D. Kuhn, C. G. Grunfelder, *Eukaryot. Cell* **2**, 769 (2003).
57. I. B. Muller, D. Domenicali-Pfister, I. Roditi, E. Vassella, *Mol. Biol. Cell* **13**, 3787 (2002).
58. A. E. Fujimura, S. S. Kinoshita, V. L. Pereira-Chioccola, M. M. Rodrigues, *Infect. Immun.* **69**, 5477 (2001).
59. G. Montagna *et al.*, *Eur. J. Biochem.* **269**, 2941 (2002).
60. M. A. Ferguson, *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **352**, 1295 (1997).
61. G. D. Pollevick, J. L. Affranchino, A. C. Frasch, D. O. Sanchez, *Mol. Biochem. Parasitol.* **47**, 247 (1991).
62. R. Agusti, A. S. Couto, O. Campetella, A. C. Frasch, R. M. de Lederkremer, *Mol. Biochem. Parasitol.* **97**, 123 (1998).
63. V. Campo *et al.*, *Mol. Biochem. Parasitol.* **133**, 81 (2004).
64. C. A. Buscaglia *et al.*, *J. Biol. Chem.* **279**, 15860 (2004).
65. C. L. Allen, J. M. Kelly, *Exp. Parasitol.* **97**, 173 (2001).
66. J. A. I. Atwood III *et al.*, *Science* **309**, 473 (2005).
67. C. Yao, J. E. Donelson, M. E. Wilson, *Mol. Biochem. Parasitol.* **132**, 1 (2003).
68. G. Manning, D. B. Whyte, R. Martinez, T. Hunter, S. Sudarsanam, *Science* **298**, 1912 (2002).
69. We thank our colleagues in the *T. cruzi* Genome Network (TcGN) and the trypanosomatid research community for their continued support and encouragement. In particular, we thank the members of the Tritryp Sequencing Consortium for their help with comparative genome annotation; J. Donelson and S. Melville for a critical review of this manuscript; A. Kerhornou for his help with GPI anchor predictions; as well as O. Campetella, I. Dórso, V. Campo, G. Montagna, and F. Aguero for their help with the surface protein analyses. Funding for this project was provided by grants from the National Institute for Allergy and Infectious Diseases (NIAID) to N.M.E.-S. (AI45038), K.D.S. and P.J.M. (AI045039), and B.A. (AI45061); the M. J. Murdoch Charitable Trust to the Seattle Biomedical Research Institute; the Beijer Foundation to B.A. M.J.L. and A.C.F. are Howard Hughes Medical Institute International Research Scholars. G.C.C. is supported in part by CNPq, Brazil. We also express our gratitude to NIAID, Burroughs Wellcome Fund (BWF), the Wellcome Trust, and WHO (Tropical Disease Research) for providing funds for several TcGN and Tritryp meetings. Special thanks to C. Hertz-Fowler and P. Mooney at the Wellcome Trust Sanger Institute (WTSI) for coordinating data exchange with The Institute for Genomic Research (TIGR), and to A. Tivey and M. Aslett for loading and updating the *T. cruzi* data in GeneDB at WTSI. This Whole-Genome Shotgun project has been deposited at the DNA Databank of Japan (DDBJ), the European Molecular Biology Laboratory (EMBL), and GenBank under the project accession AAHK00000000. The version described in this paper is the first version, AAHK01000000. All data sets and genome annotations are also available through GeneDB at www.genedb.org.

Supporting Online Material

www.sciencemag.org/cgi/content/full/309/5733/409/DC1
Materials and Methods
Tables S1 to S14
References and Notes

23 March 2005; accepted 17 June 2005
10.1126/science.1112631