42. World Development Report (World Bank, Washington DC, 2005), chap. 3.
43. Organization of Pharmaceutical Producers of India and Monitor Company Group, in *Outsourcing Opportunities in Indian Pharmaceutical Industry* (Mumbai, India, 2004); available at www.indiaoppi.com/puboutsourcing2003.htm., pp. 24–25.
44. The authors are grateful to all who contributed to this paper and to The Rockefeller Foundation for financial support. The views expressed in this paper are those of the authors in their individual capacities and do not necessarily reflect those of their respective institutions. We dedicate this work to Professor Sanjaya Lall, co-author, Friend, and colleague to us all, who passed away on 18 June 2005.

Supporting Online Material
www.sciencemag.org/cgi/content/full/309/5733/401/DC1
Materials and Methods
Fig. S1

10.1126/science.1115538

RESEARCH ARTICLE

Comparative Genomics of Trypanosomatid Parasitic Protozoa

Najib M. El-Sayed,[1,2*†] Peter J. Myler,[3,4,5*†] Gaëlle Blandin,[1] Matthew Berriman,[6] Jonathan Crabtree,[1] Gautam Aggarwal,[3] Elisabet Caler,[1] Hubert Renauld,[6] Elizabeth A. Worthey,[3] Christiane Hertz-Fowler,[6] Elodie Ghedin,[1,2] Christopher Peacock,[6] Daniella C. Bartholomeu,[1] Brian J. Haas,[1] Anh-Nhi Tran,[7] Jennifer R. Wortman,[1] U. Cecilia M. Alsmark,[8] Samuel Angiuoli,[1] Atashi Anupama,[3] Jonathan Badger,[1] Frederic Bringaud,[9] Eithon Cadag,[3] Jane M. Carlton,[1] Gustavo C. Cerqueira,[1,10] Todd Creasy,[1] Arthur L. Delcher,[1] Appolinaire Djikeng,[1] T. Martin Embley,[8] Christopher Hauser,[1] Alasdair C. Ivens,[6] Sarah K. Kummerfeld,[11] Jose B. Pereira-Leal,[11] Daniel Nilsson,[7] Jeremy Peterson,[1] Steven L. Salzberg,[1] Joshua Shallom,[1] Joana C. Silva,[1] Jaideep Sundaram,[1] Scott Westenberger,[1‡] Owen White,[1] Sara E. Melville,[12] John E. Donelson,[13] Björn Andersson,[7] Kenneth D. Stuart,[3,4] Neil Hall[6†§]

A comparison of gene content and genome architecture of *Trypanosoma brucei*, *Trypanosoma cruzi*, and *Leishmania major*, three related pathogens with different life cycles and disease pathology, revealed a conserved core proteome of about 6200 genes in large syntenic polycistronic gene clusters. Many species-specific genes, especially large surface antigen families, occur at nonsyntenic chromosome-internal and subtelomeric regions. Retroelements, structural RNAs, and gene family expansion are often associated with syntenic discontinuities that—along with gene divergence, acquisition and loss, and rearrangement within the syntenic regions—have shaped the genomes of each parasite. Contrary to recent reports, our analyses reveal no evidence that these species are descended from an ancestor that contained a photosynthetic endosymbiont.

The protozoan pathogens *Leishmania major*, *Trypanosoma cruzi*, and *Trypanosoma brucei* (family *Trypanosomatidae*, order *Kinetoplastida*) collectively cause disease and death in millions of humans and countless infections in other mammals, primarily in developing countries in tropical and subtropical regions (*1*). There are no vaccines for these diseases and only a few drugs, which are inadequate because of toxicity and resistance. Although the three pathogens (referred to here as the "Tritryps") share many general characteristics, including subcellular structures such as the kinetoplast and glycosomes, each is transmitted by a different insect and has its own life-cycle features, different target tissues, and distinct disease pathogenesis in their mammalian host [box 1 in (*2*) and fig. S1]. They also use different immune evasion strategies: *L. major* alters the function of the macrophages it infects, *T. cruzi* expresses a complex variety of surface antigens from within the cells it infects, and *T. brucei* remains extracellular but circumvents the host immune response by the periodic switching of its major surface protein (*3*).

The availability of the three draft genome sequences (*4*–*6*) allows better understanding of the genetic and evolutionary bases of the shared and distinct parasitic modes and lifestyles of these pathogens. In the accompanying Research Articles, the discussion of each species reflects the current state of knowledge for each organism. Thus, the Research Article by Berriman *et al.* (*4*) emphasizes metabolism and biochemical pathways of *T. brucei*; the Research Article by Ivens *et al.* (*5*) highlights fundamental aspects of molecular biology (transcription, translation, posttranslational modification, and proteolysis) of *L. major*; and the Research Article by El-Sayed *et al.* (*6*) focuses on repeats and retroelements, DNA replication and repair, and signaling pathways of *T. cruzi*. Here, we compare gene content and genome architecture, composition, and organization of protein domains encoded by each genome and offer an analysis of the rates of gene evolution.

Core proteome. The *T. brucei*, *L. major*, and *T. cruzi* haploid genomes contain between 25 and 55 megabases (Mb) distributed over 11 to 36 (generally) diploid chromosomes, and encode about 8100, 8300, and 12,000 protein-coding genes, respectively (Table 1). An "all-versus-all" basic local alignment search tool (BlastP) comparison of the predicted protein sequences within each of the three genomes was made using a suite of algorithms designed to collapse closely related paralogous genes. In the case of *T. cruzi*, all alleles were included because of the hybrid nature of this genome (*2*, *6*). The mutual best BlastP hits between the three collapsed proteomes were grouped as clusters of orthologous genes (COGs). Iteration of this process with manual inspection and reannotation, especially of two-way COGs (i.e., those with members in only two of the Tritryps), resulted in 6158 three-way COGs, which defined the Tritryp core proteome, as well as 1014 two-way COGs (Table 1,

[1]The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, MD 20850, USA. [2]Department of Microbiology and Tropical Medicine, George Washington University, Washington, DC 20052, USA. [3]Seattle Biomedical Research Institute, 307 Westlake Avenue North, Seattle, WA 98109–2591, USA. [4]Department of Pathobiology, University of Washington, Seattle, WA 98195, USA. [5]Division of Biomedical and Health Informatics, University of Washington, Seattle, WA 98195, USA. [6]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire CB10 1SA, UK. [7]Center for Genomics and Bioinformatics, Karolinska Institutet, Berzelius väg 35, S-171 77 Stockholm, Sweden. [8]School of Biology, The Devonshire Building, University of Newcastle, Newcastle upon Tyne NE1 7RU, UK. [9]Laboratoire de Génomique Fonctionnelle des Trypanosomatides, UMR-CNRS 5162, Université Victor Segalen Bordeaux II, 33076 Bordeaux cedex, France. [10]Departamento de Bioquímica e Imunologia, Universidade Federal de Minas Gerais, CEP 31270-901, Belo Horizonte, MD, Brazil. [11]Medical Research Council Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, UK. [12]Department of Pathology, University of Cambridge, Tennis Court Road, Cambridge, CB2 1QP, UK. [13]Department of Biochemistry, University of Iowa, 4-403 Bowen Science Building, Newton Road, Iowa City, IA 52242, 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.); nhall@tigr.org (N.H.)
‡Present address: Department of Microbiology, Immunology, and Molecular Genetics, University of California, Los Angeles, CA 90095, USA.
§Present address: The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, MD 20850, USA.

Fig. 1A, and table S1). Amino acid sequence alignment of a large sample of three-way COGs reveals an average of 57% identity between *T. brucei* and *T. cruzi*, and 44% identity between *L. major* and the two other trypanosomes, reflecting expected phylogenetic relationships (*7–10*). The intracellular parasites, *L. major* and *T. cruzi*, appear to share slightly more two-way COGs than do *T. brucei* and *T. cruzi* and considerably more than do *L. major* and *T. brucei*. The remainder of each proteome is composed of species-specific members (table S1), of which *T. cruzi* (32%) and *T. brucei* (26%) have a much greater proportion than *L. major* (12%). Because the majority of the species-specific proteins appear to be members of surface antigen families, the different numbers may relate to different strategies of survival and immune evasion used in each organism. Other species-specific proteins carry out distinct metabolic and physiological functions, some of which are discussed below [see also (*4–6*)].

Species-specific protein domains. A comparison of the Pfam and TIGRFAMs protein domains in the Tritryp genomes revealed very few that are unique to individual organisms (Fig. 1B). Of the 1617 protein domains identified in the Tritryp genomes, fewer than 5% are unique to a single species (table S2). For example, macrophage migration inhibitory factor (Pfam accession number PF01187) domain, restricted to *L. major*, may inhibit macrophage activation and consequent destruction of the parasites, as described in *Brugia malayi* (*5*, *11*). Another domain specific to *L. major* (PF04133) is involved in vacuolar transport, suggesting that the protein may act to divert proteases within the host phagolysosome. These domains are not seen in *T. cruzi*, which escapes from the lysosomal compartment into the cytoplasm soon after invasion, or in *T. brucei*, which is extracellular.

The variant surface glycoprotein (VSG) expression site–associated gene (ESAG) domains ESAG1 (PF03238) and ESAG6-7 (PF05446) are restricted to *T. brucei*. Likewise, the AOX domain (PF01786), which acts as an alternative terminal oxidase in mitochondria, and the LigB domain (PF02900), which is involved in aromatic compound metabolism, account for some of the few metabolic capabilities of *T. brucei* that are not found in *L. major* or *T. cruzi* (*4–6*).

T. cruzi has a serine carboxypeptidase S28 domain (PF05577) not found in *T. brucei* or *L. major*. Several lines of evidence indicate that *T. cruzi* secretes a small peptide processed by a serine peptidase, which interacts with a host-cell receptor in a wide variety of mammalian cells (*12*). This interaction leads to a calcium signaling reaction that triggers lysosome migration to the host-cell plasma membrane, enabling parasite entry (*13*). *T. cruzi* also contains a number of hormone-type domains such as PF00220 (neurohypophysial hormones, N-terminal domain) and the PF02044 (bombesin-like peptide), which are not found in *T. brucei* or *L. major*, but the functional significance of these domains is uncertain.

Specific domain expansion and loss. Several interesting examples of domain expansion or contraction (table S3) were revealed (fig. S2), similar to those seen in other parasites such as *Plasmodium* (*14*). Many of these proteins appear to be involved in host interactions and often are encoded in tandem arrays, typically at species-specific subtelomeric locations (*4–6*). For example, *T. brucei* has expanded ESAG4 proteins that contain adenylate and guanylate cyclase catalytic domains and proteins containing leucine-rich repeat domains. *T. cruzi* has expanded bacterial neuraminidase/Asp-box repeat, mucin-like glycoprotein, leishmanolysin, and trypanosome retrotransposon hot spot (RHS) domains in trans-sialidases, mucins, glycoprotein (gp) 63 proteases, and RHS proteins, respectively. *L. major* contains a large tandem array of amastin surface glycoproteins but also possesses expanded protein families containing mitochondrial carrier protein, adenosine 5′-triphosphate–binding cassette transporters, and heat shock protein (HSP) 90. Interestingly, compared with *T. brucei* and *T. cruzi*, *L. major* has a marked underrepresentation of domains involved in RNA binding (pumilio and zinc finger domains), protein-protein interaction (leucine-rich and tetratricopeptide repeats), and calcium signaling (calmodulin and EF hand), suggesting a reduced role or alternate pathways for these activities.

Large-scale synteny. Despite having diverged 200 to 500 million years ago (*15–18*) and thus predating the emergence of mammals (*19*), the genomes of the trypanosomatid species are highly syntenic (i.e., show conservation of gene order). Of all the genes in *T. brucei* and *L. major*, 68 and 75%, respectively, remain in the same genomic context. Moreover, almost all (94%) of the three-way COGs that form the core proteome fall within regions of conserved synteny. The *T. brucei* and *L. major* genomes (*2*) show 110 blocks of synteny spanning 19.9 and 30.7 Mb, respectively (Fig. 2). Detailed examination of the synteny breakpoints revealed that 40% were associated with expansions of multigene families, retroelements and/or structural RNAs (Plate 1, fig.

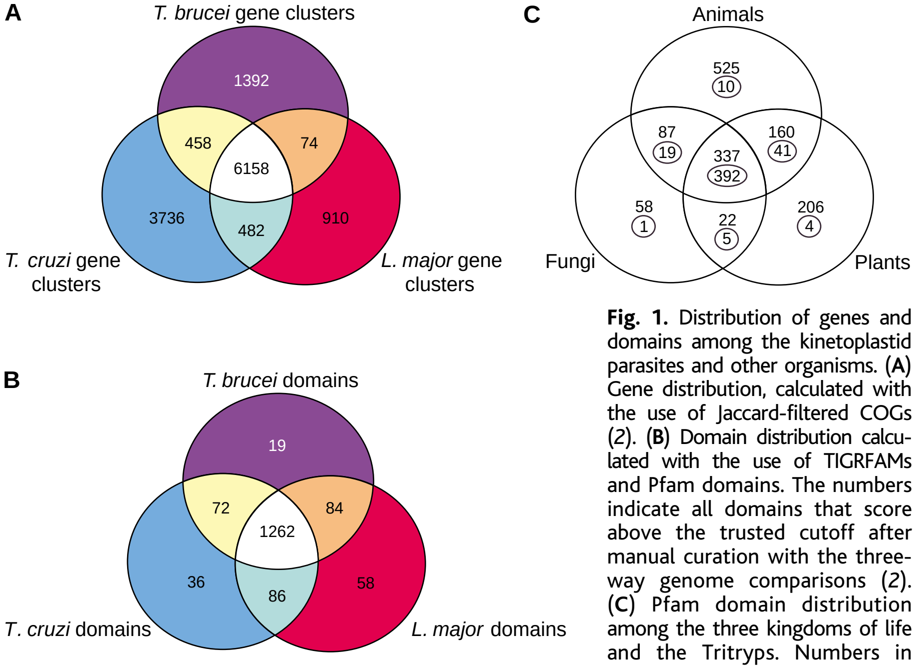

Fig. 1. Distribution of genes and domains among the kinetoplastid parasites and other organisms. **(A)** Gene distribution, calculated with the use of Jaccard-filtered COGs (*2*). **(B)** Domain distribution calculated with the use of TIGRFAMs and Pfam domains. The numbers indicate all domains that score above the trusted cutoff after manual curation with the three-way genome comparisons (*2*). **(C)** Pfam domain distribution among the three kingdoms of life and the Tritryps. Numbers in small circles indicate the number of domains that occur more than once in Tritryp parasite genomes. The numbers above the small circles indicate Pfam domains that are not present in the Tritryps.

Table 1. General features of the Tritryp genomes. We found 5812 syntenic three-way COGs and 346 nonsyntenic three-way COGs. Mbp, mega–base pairs; NC, not computed.

	T. brucei	*T. cruzi*	*L. major*
Haploid genome size (Mbp)	25*	55	33
No. of chromosomes (per haploid genome)	11*	~28†	36
No. of genes (per haploid genome)	9068‡	~12,000§	8311‖
Total regions with synteny blocks (Mbp)	19.9	NC	30.7
Mean CDS size (bp) in syntenic three-way COGs	1511	1457	1731
Mean inter-CDS size (bp) between syntenic three-way COGs	721	561	1431

*Excluding ~100 mini- and intermediate-sized chromosomes (totaling ~10 Mb). †The exact number is not known and homologs can differ substantially in size. ‡Includes 904 pseudogenes. §The exact number of haploid genes has not been determined in *T. cruzi*. ‖Includes 34 pseudogenes.

S3, and table S4). Enrichment of segmental duplication in regions of synteny breakpoints has also been observed in mammals (*20*, *21*), but the implications are unknown. Interestingly, 43% of the synteny breakpoints in *T. brucei* and *L. major* (excluding chromosome ends) occur at or very close to the strand-switch regions separating the directional gene clusters (DGCs), which are characteristic of and unique to trypanosomatid genomes (*5*). Thus, there appears to be strong selective pressure to maintain gene order and to keep the DGCs intact, despite the extensive sequence divergence between the genes themselves. This may also be related to the relatively low incidence of sexual recombination in these organisms (*22*), which would limit opportunities for rearrangement during meiosis.

Localized chromosomal rearrangements. Despite the marked overall conservation, many local insertions, deletions, or substitutions were seen within otherwise syntenic regions. Although evidence for all three processes was found in the Tritryp genomes, gene insertions or substitutions (which result in species-specific genes and nonsyntenic three-way COGs) were more common than gene loss (two-way COGs that include a *L. major* gene). Some of these events can result in substantial physiological and biochemical differences between these parasites.

One example of an insertion involves two genes in *T. brucei* encoding subunits (ESAG6 and ESAG7) of the heterodimeric transferrin receptor (*23*) in the syntenic block Tb7.3/Lm 22.1 (Plate 1, inset A). The two genes are more similar (97% amino acid identity) to one another than they are to any of the subtelomeric copies (73 to 77% identity), indicating they encode a different form of the receptor than the telomeric copies. The surrounding region in this synteny block contains other insertions specific to *L. major* and *T. cruzi* and several translocated genes in the three genomes. *T. cruzi* seems to have undergone four separate insertions of genes belonging to a metabolic pathway that converts L-histidine to L-glutamate (*5*). Intriguingly, the gene (*hutG*) for the final enzyme has been previously found only in bacteria, and this may mark a horizontal gene transfer from bacteria, where the genes occur in a single operon.

Interestingly, a component of the RNA interference (RNAi) pathway is present only in *T. brucei* (Tb10.406.0020/*TbAGO1*) (*24, 25*) and a gene (LmjF33.0290) encoding a glucose transporter (*26*) is present only in *L. major* within the same synteny block (Plate 1, inset B). This region is also associated with a cluster of tRNA

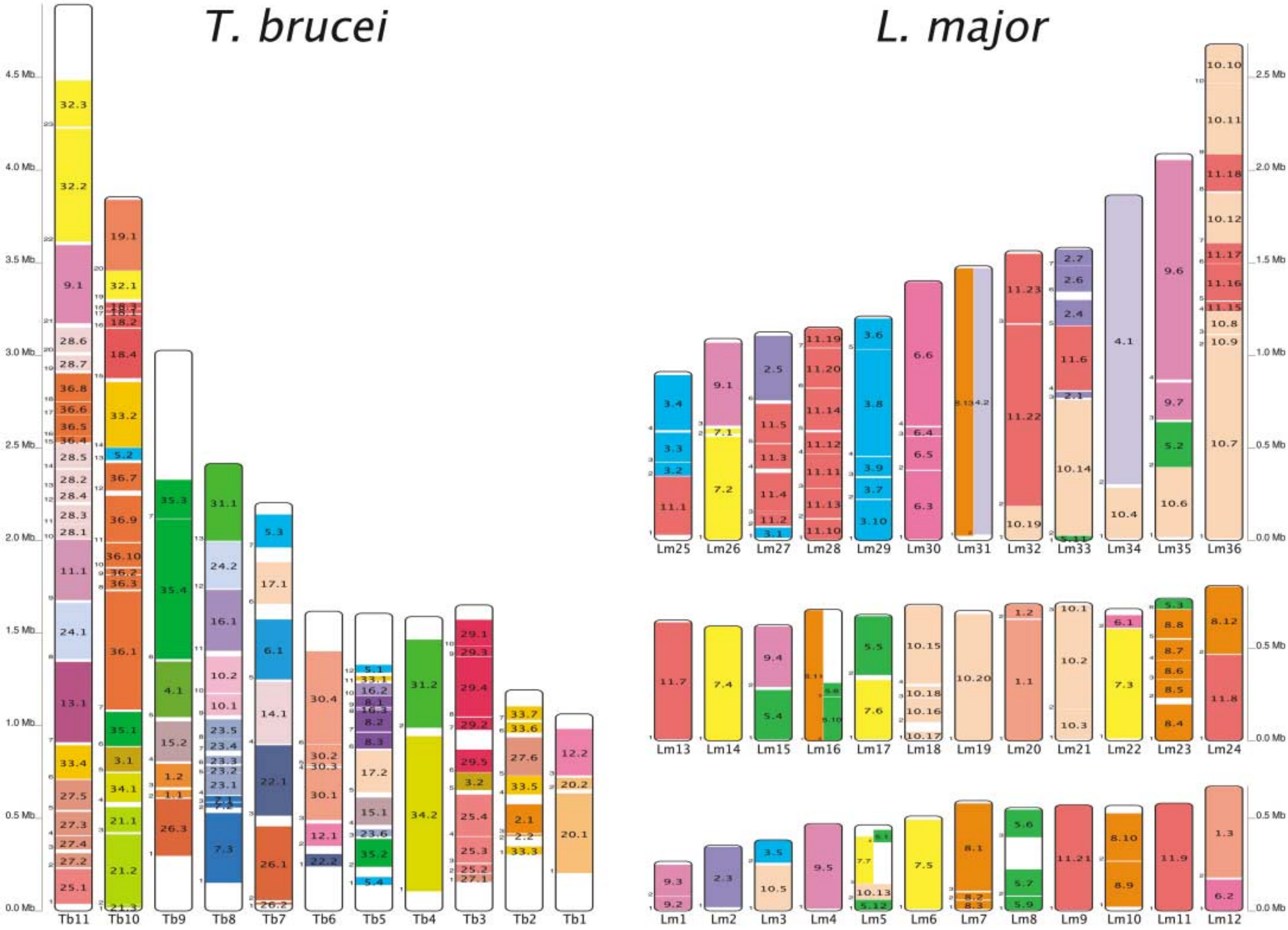

Fig. 2. Synteny maps. The 36 different colors in the *T. brucei* (left) panel represent the locations of the indicated synteny blocks in the 36 chromosomes of *L. major*, and the 11 colors in the *L. major* (right) panel depict the locations of the indicated synteny blocks in the 11 chromosomes of *T. brucei*. Each synteny block is named using a double nomenclature that refers to the chromosomal location of the block in both species. Labels on the left outside margin of the synteny blocks denote the block number in the reference genome. Labels within synteny blocks refer to their location on the other genome. For example, synteny block Tb1.1 of *T. brucei* chromosome 1 (Tb1; lower right of left panel) is synteny block Lm20.1 of *L. major* chromosome 20 (Lm20). As another example, all of the yellow synteny blocks in the *L. major* panel (blocks Tb7.1 to Tb7.7) are on *T. brucei* chromosome 7 (Tb7). Synteny blocks are defined as groups of five or more *T. brucei* genes that possess an ortholog on the same *L. major* chromosome (*2*). The entire map contains 7974 *T. brucei* and 7466 *L. major* protein-coding genes in 110 synteny blocks. Plate 1 and fig. S3, A to K, show more detailed views and table S4, A and B, has complete lists of genes and block coordinates.

genes and *HSP83* arrays (*27*, *28*) of variable lengths in the three genomes. In eukaryotes, RNAi maintains genome integrity and prevents invasion by nucleic acid, by means of double-stranded RNAs that induce the degradation of homologous mRNAs (*25*). None of the Tritryps possesses an obvious homolog of *Dicer*, an essential component of the RNAi pathway in other organisms, but two predicted proteins (named TbRN3A and TbRN3B; table S5) containing a single ribonuclease III domain are present in *T. brucei* (but are not seen in *L. major* or *T. cruzi*) and represent potential *Dicer* candidates. Thus, gene insertions are likely responsible for the RNAi activity seen in *T. brucei* but not *T. cruzi* or *L. major* (*28*), although the source(s) of the gene insertions remains unknown.

Both *L. major* and *T. cruzi* contain a gene encoding 5-oxoprolinase in synteny block Tb10.15/Lm18.4, which is absent from *T. brucei* (Plate 1, inset C) and thus appears to represent gene deletion. This enzyme catalyzes the formation of L-glutamate from 5-oxo-L-proline in the glutathione metabolism pathway. Further examples of small synteny breaks occur immediately upstream of this same region: *T. brucei* contains a receptor-type adenylate cyclase gene (GRESAG4) gene; *T. cruzi* contains one or two dispersed gene family 1 (DGF-1) pseudogenes, as well as a RHS pseudogene on one allele; and *L. major* contains a degenerate *Ingi*/L1*Tc*-related retroelement (DIRE). These sequences are often associated with frequent recombination in the Tritryps (*6*) and large synteny breaks.

Nonsyntenic and subtelomeric regions. Plotting the location of two-way and three-way COGs across the *L. major* and *T. brucei* genomes revealed that synteny extends over most of both genomes (Fig. 2 and fig. S4). *L. major* contains a few large chromosome-internal nonsyntenic regions, which mostly correspond to tandem arrays of both protein-coding and RNA genes. By contrast, *T. brucei* contains large blocks of nonsyntenic genes at the telomeres of all chromosomes (fig. S4), which can be several hundred kilobases (kb) in size and contain large arrays of species-specific VSG (pseudo)genes and ESAGs, as well as a large number of retroelements and RHS genes (*4*). The subtelomeric regions of *T. cruzi* are also large and nonsyntenic, consisting mostly of interspersed arrays of trans-sialidase superfamily, DGF-1 and RHS (pseudo)-genes, as well as vestigial interposed retroelements (VIPER), short interspersed repetitive elements (SIRE), *T. cruzi* L1*Tc* *T. cruzi* nonautonomous non-LTR retrotransposons (NAR*Tc*), and/or DIRE retroelements (*6*). Another intriguing feature of *T. cruzi* is the presence of large (up to 600 kb) nonsyntenic "islands" of genes coding for surface proteins such as trans-sialidase, mucin, mucin-associated surface protein (MASP), and gp63 peptidase, along with retrotransposons and RHS genes. Although the precise location of these "islands" is not certain, they appear often to lie between chromosome-internal synteny blocks. The *L. major* subtelomeric regions are quite short (<20 kb), with relatively few repetitive sequences, although there is evidence that there may be some recombination between telomeres (*5*). Nevertheless, the most telomere-proximal genes in *L. major* are often nonsyntenic, although usually not specific to *L. major*. Thus, the organization and gene content of the subtelomeric regions is quite different in each genome (Fig. 3).

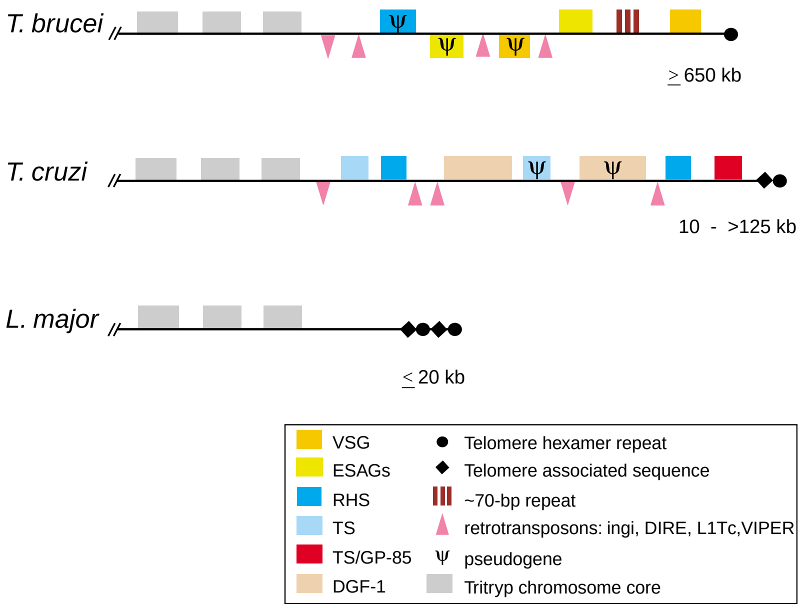

Fig. 3. Prototypes of Tritryp subtelomeric regions. Subtelomeric regions are defined here as the area that extends from the telomeric hexamer repeats to the first nonrepetitive sequence. Boxes indicate genes and/or gene arrays. Genes and/or gene arrays shown above the line are oriented toward the telomeres, whereas those shown below the line are oriented in the opposite direction. The size range of the subtelomeric regions in each genome is indicated on the right. The TS and TS/GP-85 boxes depict the trans-sialidase and GP-85 trans-sialidase superfamilies, respectively.

Chromosome evolution. A comparison of the Tritryp genomes provides interesting insights into the karyotype of their common ancestor. *T. brucei* has only 11 large diploid chromosomes (plus numerous small chromosomes that contain largely repetitive sequence), and *T. cruzi* and *L. major* contain ~28 and 36 pairs of smaller chromosomes, respectively. Most rearrangements of synteny blocks represent inversions and/or translocations (Fig. 2 and fig. S3), but there appear to be several cases of chromosome fusions in *T. brucei*. Twenty of the 36 *L. major* chromosomes are almost entirely syntenic within a substantially larger *T. brucei* chromosome, except for a few instances of synteny block inversion or shuffling. In 10 further cases, there is only a single segmental translocation that has moved one end of the *L. major* chromosome to a different *T. brucei* chromosome. Although the *T. cruzi* genome contains gaps, many nearly chromosome-sized scaffolds were defined by virtue of arrays of telomeric repeats and characteristic subtelomeric genes at one end (*6*). Notably, many of the *L. major* chromosomes that are syntenic with these *T. cruzi* scaffolds also contain telomeric sequences at the corresponding end. In contrast, the syntenic *T. brucei* regions at the corresponding position generally represent internal chromosome regions with no typical telomeric structures. For example, the ends of the two *L. major* and *T. cruzi* chromosomes appear to have joined to form a single *T. brucei* chromosome at the junction between synteny blocks Tb11.7/Lm13.1 and Tb11.8/Lm24.1 (Plate 1). Interestingly, this synteny break region in *T. brucei* contains RHS, DIRE, and *Ingi* sequences, often associated with *T. brucei* subtelomeres, pointing to a telomeric origin for this region. Other examples of similar apparent chromosome fusions in *T. brucei* can be seen between synteny blocks Tb7.3/Lm22.1 and Tb7.4/Lm14.1, as well as Tb7.5/Lm6.1 and Tb7.6/Lm.1. Thus, the current chromosomal architecture of *T. brucei* seems to have derived from an ancestor with the more fragmented genomic organization of *L. major* and *T. cruzi*. This evolutionary topology supports the prevailing view of an early divergence of the *Leishmania* genus and the monophyly of the *Trypanosoma* genus (*7*–*10*).

The marked difference in the gene size and density between the Tritryp genomes is notable. The average *L. major* protein-coding sequence (CDS) is considerably longer than in *T. brucei* or *T. cruzi* (Table 1), often in regions specifying low-complexity amino acid insertions or expansions. The length differential is even more extreme in noncoding regions, with the average inter-CDS length in *L. major* being almost twice that in *T. brucei* and three times that in *T. cruzi* (Table 1). Consequently, gene density in *L. major* is considerably less than in *T. brucei* and *T. cruzi* (251 versus 319 and 385 genes/Mb, respectively). Thus, genome compaction does not appear to be associated with an intracellular lifestyle in the Tritryps, in contrast to the suggestion for *Encephalitozoon cuniculi* (*29*, *30*).

The Tritryp chromosomes exhibited systematic purine excess, GC bias, and AT skew, correlated with the coding strand (*31*). This phenomenon is associated with replication in Eubacteria and Archaea (*32*), but local skews are linked to mutational bias arising from transcription in eukaryotes (*33*). Although this seems to be the case for *L. major*, GC skew has the opposite correlation with the coding strand in *T. brucei* and *T. cruzi* (*34*). The AT skew correlation is the same in all three species. Thus, there may be differences in DNA repair and/or transcriptional processes between *Leishmania* and *Trypanosoma*, which may account for their different GC contents (60% in *L. major*, 46 to 51% in *T. brucei* and *T. cruzi*).

No evidence for ancestral plastid endosymbiont. On the basis of recently reported evidence of plantlike traits associated with the metabolism of *Trypanosoma* parasites (*35*) and the close phylogenetic relatedness of kinetoplastids to *Euglena* (a photosynthetic protist), it has been suggested that the common ancestor of these protists harbored the same endosymbiotic green alga that gave rise to the secondary plastid in *Euglena*. Our data show that the protein domain content of the Tritryps is not consistent with large-scale horizontal transfer of genetic material from plants, given that we did not observe a large number of plant-specific domains in the Tritryps (Fig. 1C and table S6).

We used phylogenetic analyses to search for genes of cyanobacterial or green algal ancestry in the Tritryp genomes. Phylogenetic trees were made for all *L. major* genes (because this genome has the fewest protein-coding genes) with the use of alignments against proteins from all available completed genomes (*2*). Although some genes appeared to branch with plants or cyanobacteria in an initial screen, these relationships were not supported by more sophisticated Bayesian methods and a more comprehensive sampling of protein sequences (*2*). These analyses included the genes previously reported to have plantlike traits (*35*), as shown in fig. S5. We conclude from our analyses that the genome data provide no unambiguous support for the hypothesis that trypanosomatids have acquired genes from the endosymbiont that gave rise to the Euglena secondary plastid, suggesting that it was acquired subsequent to speciation with the kinetoplastida.

Gene evolution. Pathogen proteins involved in interaction with the host are often rapidly evolving, and can be identified by comparison of the number of synonymous mutations per synonymous site (d_S) and the number of nonsynonymous mutations per nonsynonymous site (d_N) (*36*). As the Tritryp genomes are too divergent to accurately estimate d_S, we calculated d_N using pairwise comparisons for every COG where there was a simple 1:1:1 orthologous relationship between genomes (or 1:1:2 in cases where both *T. cruzi* alleles were present), because this effectively gives a measure of how rapidly each protein sequence is diverging between species (*2*). Categorization of these genes by gene ontology (GO) term for biological processes (Fig. 4) showed that those with no functional annotation had the highest median d_N value, suggesting that they were subject to positive selection causing active accumulation of mutations or that they were under neutral evolution allowing the sequences to drift. Such genes of unknown function probably include trypanosomatid-specific genes involved in unique processes (including interaction with the host) or highly variable genes that elude annotation by homology.

Genes in the transport category also had a relatively high median d_N value for *L. major* versus *T. brucei* (Fig. 4). Rapid evolution of transport proteins may be due to their surface location (and consequent exposure to the host immune system) but may also reflect the different niches occupied by each parasite within their hosts and requirement for different nutrient uptake from their environment. Conversely, genes representing metabolism, cell growth, and maintenance have low d_N values, probably reflecting the core processes common to the Tritryps.

Concluding remarks. Although the majority of trypanosomatid genes in the same genomic context are conserved, there are substantial differences, which presumably reflect specific adaptations to distinct species-specific selection pressures and the distinct pathophysiologies and survival strategies of each organism. Antigenic variation and diversity are characteristic of *T. brucei* and *T. cruzi*, and the localization of large arrays of genes encoding surface proteins at or near telomeres and/or the presence of numerous retroelements within these regions may enhance recombination frequency and provide for rapid sequence variation. Thus, colocalization of previously uncharacterized genes (e.g., *T. cruzi* MASPs and DGF-1, as well as RHS in both *T. cruzi* and *T. brucei*) in these regions leads to the suspicion that they may also be involved with immune evasion or survival in different hosts. The frequent recombination in these regions results in large (up to 2 Mb) size polymorphisms between homologous chromosomes seen in *T. brucei* and *T. cruzi*.

The frequent correlation between conserved synteny blocks and the large DGCs characteristic of the Tritryps may also reflect their unique linkage of transcription with subsequent RNA processing by trans-splicing and polyadenylation. Transcription of protein-coding genes has been postulated to initiate at only a few sites on each chromosome (*37*–*39*), suggesting that there may be selective pressure against synteny breaks within the polycistronic gene clusters downstream of these sites. It is also possible (and not necessarily unrelated) that the synteny breaks associated with strand-switch regions may reflect higher rates of recombination at these sites, possibly as a result of linkage with replication processes.

The identification of numerous Tritryp-conserved and species-specific genes provides the opportunity for development of previously unexplored chemotherapeutic approaches against these parasites. Drugs designed against conserved core processes hold the advantage of being potentially useful against all three organisms, provided that they are sufficiently divergent from mammalian host proteins.

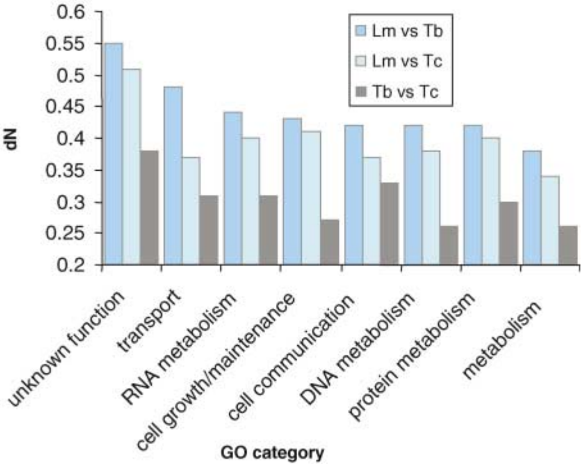

Fig. 4. Median d_N values for genes categorized by GO process annotation. Each bar corresponds to a pairwise comparison of sets of orthologous genes from two species. Amino acid sequences were aligned for each gene and converted to nucleotide sequences to calculate nonsynonymous substitutions (*40*). Genes were annotated with GO process terms in *T. brucei* and then transferred to the other species with the use of orthology. Results from the selective constraint analyses are in table S7, A to C.

References and Notes

1. M. P. Barrett *et al.*, *Lancet* **362**, 1469 (2003).
2. Materials and methods are available as supporting material on *Science* Online.
3. E. Pays, L. Vanhamme, D. Perez-Morga, *Curr. Opin. Microbiol.* **7**, 369 (2004).
4. M. Berriman *et al.*, *Science* **309**, 416 (2005).
5. A. C. Ivens *et al.*, *Science* **309**, 436 (2005).
6. N. M. El-Sayed *et al.*, *Science* **309**, 409 (2005).
7. J. Haag, C. O'hUigin, P. Overath, *Mol. Biochem. Parasitol.* **91**, 37 (1998).
8. J. Lukes *et al.*, *J. Mol. Evol.* **44**, 521 (1997).
9. A. D. Wright, S. Li, S. Feng, D. S. Martin, D. H. Lynn, *Mol. Biochem. Parasitol.* **99**, 69 (1999).
10. J. R. Stevens, H. A. Noyes, G. A. Dover, W. C. Gibson, *Parasitology* **118**, 107 (1999).
11. F. H. Falcone *et al.*, *J. Immunol.* **167**, 5348 (2001).
12. B. A. Burleigh, N. W. Andrews, *Curr. Opin. Microbiol.* **1**, 461 (1998).

13. E. V. Caler, S. Vaena de Avalos, P. A. Haynes, N. W. Andrews, B. A. Burleigh, *EMBO J.* **17**, 4975 (1998).
14. J. Carlton, J. Silva, N. Hall, *Curr. Issues Mol. Biol.* **7**, 23 (2005).
15. P. Overath, J. Haag, A. Lischke, C. O'hUigin, *Int. J. Parasitol.* **31**, 468 (2001).
16. J. R. Stevens, H. A. Noyes, C. J. Schofield, W. Gibson, *Adv. Parasitol.* **48**, 1 (2001).
17. J. R. Stevens, W. C. Gibson, *Cad. Saude Publica* **15**, 673 (1999).
18. E. J. Douzery, E. A. Snell, E. Bapteste, F. Delsuc, H. Philippe, *Proc. Natl. Acad. Sci. U.S.A.* **101**, 15386 (2004).
19. S. J. O'Brien *et al.*, *Science* **286**, 458 (1999).
20. J. A. Bailey, R. Baertsch, W. J. Kent, D. Haussler, E. E. Eichler, *Genome Biol.* **5**, R23 (2004).
21. L. Armengol, M. A. Pujana, J. Cheung, S. W. Scherer, X. Estivill, *Hum. Mol. Genet.* **12**, 2201 (2003).
22. W. Gibson, J. Stevens, *Adv. Parasitol.* **43**, 1 (1999).
23. D. Salmon *et al.*, *Cell* **78**, 75 (1994).
24. H. Shi, A. Djikeng, C. Tschudi, E. Ullu, *Mol. Cell. Biol.* **24**, 420 (2004).
25. E. Ullu, C. Tschudi, T. Chakraborty, *Cell. Microbiol.* **6**, 509 (2004).
26. C. K. Langford, S. A. Ewbank, S. S. Hanson, B. Ullman, S. M. Landfear, *Mol. Biochem. Parasitol.* **55**, 51 (1992).
27. E. A. Dragon, S. R. Sias, E. A. Kato, J. D. Gabe, *Mol. Cell. Biol.* **7**, 1271 (1987).
28. J. C. Mottram, W. J. Murphy, N. Agabian, *Mol. Biochem. Parasitol.* **37**, 115 (1989).
29. M. D. Katinka *et al.*, *Nature* **414**, 450 (2001).
30. J. Zhang, *Trends Genet.* **16**, 107 (2000).
31. P. D. McDonagh, P. J. Myler, K. Stuart, *Nucleic Acids Res.* **28**, 2800 (2000).
32. J. R. Lobry, *Mol. Biol. Evol.* **13**, 660 (1996).
33. S. Aerts, G. Thijs, M. Dabrowski, Y. Moreau, B. De Moor, *BMC Genomics* **5**, 34 (2004).
34. D. Nilsson, B. Andersson, *Exp. Parasitol.* **109**, 143 (2005).
35. V. Hannaert *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **100**, 1067 (2003).
36. N. Hall *et al.*, *Science* **307**, 82 (2005).
37. E. A. Worthey *et al.*, *Nucleic Acids Res.* **31**, 4201 (2003).
38. S. Martinez-Calvillo *et al.*, *Mol. Cell* **11**, 1291 (2003).
39. C. E. Clayton, *EMBO J.* **21**, 1881 (2002).
40. H. Higo *et al.*, *Int. J. Parasitol.* **27**, 1369 (1997).

41. Funding for this project was provided by grants from National Institute for Allergy and Infectious Disease (NIAID) to N.M.E.-S. (AI45038), K.S. and P.J.M. (AI045039), and B.A. (AI45061) and the Wellcome Trust (WT) to the Sanger Institute. We also thank NIAID, the Burroughs Wellcome Fund (BWF), WT, and the World Health Organization Special Programme for Research and Training in Tropical Diseases (WHO TDR) for providing funds for several Tritryp meetings. G.C.C. is the recipient of a fellowship from the Conselho Nacional de Desenvolvimento Científico e Technológico (CNPq-Brazil). *L. major* genome accession numbers: AE001274, CP000078 to CP000081, NC_004916, AL389894, AL139794, and consecutive accession numbers CT005244 to CT005272. *T. brucei* genome accession numbers: Sequence data have been deposited at DDBJ/EMBL/GenBank with consecutive accession numbers CP000066 to CP000071 for chromosomes 3 to 8 and project accession numbers AAGZ00000000, AAHA00000000, 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 have accession number CR940345. *T. cruzi* genome accession numbers: This Whole-Genome Shotgun project has been deposited at DDBJ/EMBL/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/309/5733/404/DC1
Materials and Methods
Figs. S1 to S5
Tables S1 to S7
References

14 March 2005; accepted 21 June 2005
10.1126/science.1112181

RESEARCH ARTICLE

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