SCIENTIFIC DATA

OPEN

SUBJECT CATEGORIES

» Evolutionary biology

» Parasite evolution

» DNA sequencing

» Genome

A draft genome for the African crocodilian trypanosome *Trypanosoma grayi*

Steven Kelly[1], Alasdair Ivens[2], Paul T. Manna[3], Wendy Gibson[4] & Mark C. Field[3]

Received: 16 May 2014

Accepted: 09 July 2014

Published: 05 August 2014

The availability of genome sequence data has greatly enhanced our understanding of the adaptations of trypanosomatid parasites to their respective host environments. However, these studies remain somewhat restricted by modest taxon sampling, generally due to focus on the most important pathogens of humans. To address this problem, at least in part, we are releasing a draft genome sequence for the African crocodilian trypanosome, *Trypanosoma grayi* ANR4. This dataset comprises genomic DNA sequences assembled *de novo* into contigs, encompassing over 10,000 annotated putative open reading frames and predicted protein products. Using phylogenomic approaches we demonstrate that *T. grayi* is more closely related to *Trypanosoma cruzi* than it is to the African trypanosomes *T. brucei*, *T. congolense* and *T. vivax*, despite the fact *T. grayi* and the African trypanosomes are each transmitted by tsetse flies. The data are deposited in publicly accessible repositories where we hope they will prove useful to the community in evolutionary studies of the trypanosomatids.

Design Type(s)	genome sequencing • observation design
Measurement Type(s)	genome sequencing
Technology Type(s)	next generation sequencing
Factor Type(s)	
Sample Characteristic(s)	*Trypanosoma grayi*

[1]Department of Plant Sciences, University of Oxford, Oxford OX1 3RB, UK. [2]Centre for Immunity, Infection and Evolution, Ashworth Laboratories, University of Edinburgh, Edinburgh EH9 3JT, UK. [3]Division of Biological Chemistry and Drug Discovery, University of Dundee, Dundee DD1 5EH, Scotland, UK. [4]School of Biological Sciences, University of Bristol, Bristol BS8 1UG, UK.
Correspondence and requests for materials should be addressed to M.C.F. (m.c.field@dundee.ac.uk)

Background & summary

Most trypanosomatid parasites of humans, livestock and plants are transmitted between hosts by invertebrate vectors. They are widespread, and are collectively responsible for substantial economic and health losses in many of the world's poorest regions. Within this group are the *Leishmania* species, the causative agents of Leishmaniasis, as well as the monophyletic *Trypanosoma* genus, which includes *Trypanosoma cruzi* and *Trypanosoma brucei*, causative organisms of Chagas disease and African trypanosomiasis respectively. Despite a probable single origin of vertebrate parasitism within this monophyletic group[1,2], the challenge of escaping clearance by host immune responses has given rise to a variety of disparate parasitic lifestyles including intracellular parasitism (*Leishmania spp.* and *T. cruzi*) and antigenic variation (*T. brucei*)[3–5].

Since the publication of the *T. cruzi*, *T. brucei* and *Leishmania major* genomes in 2005[6–8], several other trypanosomatid genome sequences have been released, including further *Leishmania* species and other African tsetse transmitted trypanosomes related to *T. brucei*[9–12]. The availability of sequence data for other *T. brucei* clade trypanosomes has increased our understanding of the evolution of the primary immune evasion strategy of this parasite as well as the evolution of cell surface molecules that represent the host-parasite interface[12,13]. Similarly, sequencing of disparate *T. cruzi* isolates has provided major insights into population structure and dynamics[14,15]. Though more species now have published genome sequences, sampling across the trypanosomatid phylogeny is limited and thus there are limited resources for comparative genomic investigations.

To address this key knowledge gap, here we provide a draft genome sequence of the African trypanosomatid parasite of crocodiles, *Trypanosoma grayi* (Data Citation 1 and Data Citation 2). *T. grayi* is an extracellular parasite of the bloodstream of crocodiles, and though it is transmitted by tsetse flies it is closely related to other trypanosome parasites of crocodiles in South America[16]. The trypanosome is taken up by tsetse flies in a bloodmeal and resides solely within the mid- and hindgut. Unlike salivarian trypanosomes, transmission between crocodile hosts occurs *via* oral contamination with infective metacyclics in tsetse faeces[17,18]. This faecal transmission strategy is employed by many other trypanosomes, including *T. cruzi*.

BLAST and OrthoMCL analysis of the genome sequence and predicted gene models respectively suggests that *T. grayi* possesses neither the *T. brucei* type VSG surface antigens nor the *T. cruzi* type mucin coat. Thus *T. grayi* may have evolved an alternative family of primary surface antigen genes, or possess a novel immune evasion strategy geared to survival in the reptilian bloodstream[19]. Both phylogenomic reconstruction and best-BLASTp analysis demonstrate that *T. grayi* is more closely related to *T. cruzi* than to *T. brucei* (Figure 1 and Table 1). This result refines the phylogenetic position of *T. grayi,* that in previous studies using 18S ribosomal RNA and glycosomal glyceraldehyde dehydrogense (gGAPDH) genes was placed in a separate clade from both *T. cruzi* and *T. brucei*, often with other reptile or bird trypanosomes[20–22]. Additional taxon sampling in this region of the phylogenetic tree will be important for resolving these relationships further. We anticipate that these data will provide a useful comparator for evolutionary studies of the adaptations of trypanosomes to different vertebrate hosts, as well as increasing the available sequence data resources for this globally important group of parasites.

To generate the draft genome, DNA from *T. grayi* strain ANR4, isolated from the midgut of the tsetse fly *Glossina palpalis gambiensis* in The Gambia[20] was sequenced by 91 bp paired-end Illumina sequencing and assembled *de novo* into contigs (Data Citation 2). We inferred the phylogenetic position of *T. grayi* strain ANR4 through construction of a concatenated protein sequence phylogeny using 959 single copy nuclear encoded genes. We also confirmed that both the 18S ribosomal RNA sequence and gGAPDH sequence for our *T. grayi* strain ANR4 were 100% identical to those provided in GenBank (AJ005278 and AJ620257 respectively) for *T. grayi*. Furthermore, we have identified and annotated over 10,000 putative open reading frames and have submitted this information to public databases alongside the draft genome sequence.

Methods

Sequencing and assembly

T. grayi strain ANR4 was grown *in vitro* in Cunningham's medium and genomic DNA was extracted from agarose plugs using standard phenol/chloroform methods. DNA was sequenced by 91 bp paired-end Illumina sequencing at the Beijing Genomics institute (www.genomics.cn/en/). Raw reads were subject to quality filtering using trimmomatic[23]. This was done to remove low quality bases and read-pairs as well as contaminating adaptor sequences prior to assembly. Sequences were searched for all common Illumina adaptors (the default option) and the settings used for read processing by trimmomatic were 'LEADING:10 TRAILING:10 SLIDINGWINDOW:5:15 MINLEN:50'. The quality filtered paired-end reads were then subject to read error correction using the ALLPATHS-LG[24] ErrorCorreactReads.pl program using the default program settings. The corrected reads were then assembled using SGA[25] using default settings and setting the mimimum overlap length to 80. The assembled contigs were scaffolded by mapping the trimmed and filtered paired-end reads (described above) to the assembled contigs using BWA-MEM and scaffolding the contigs using the SGA[25] scaffolding algorithm using default program settings. The resultant scaffolds were then subject to fourteen rounds of assembly error correction and gap filling using Pilon (http://www.broadinstitute.org/software/pilon/) using the '–fix all' option and setting the expected ploidy to diploid. Following scaffolding and assembly error correction all filtered

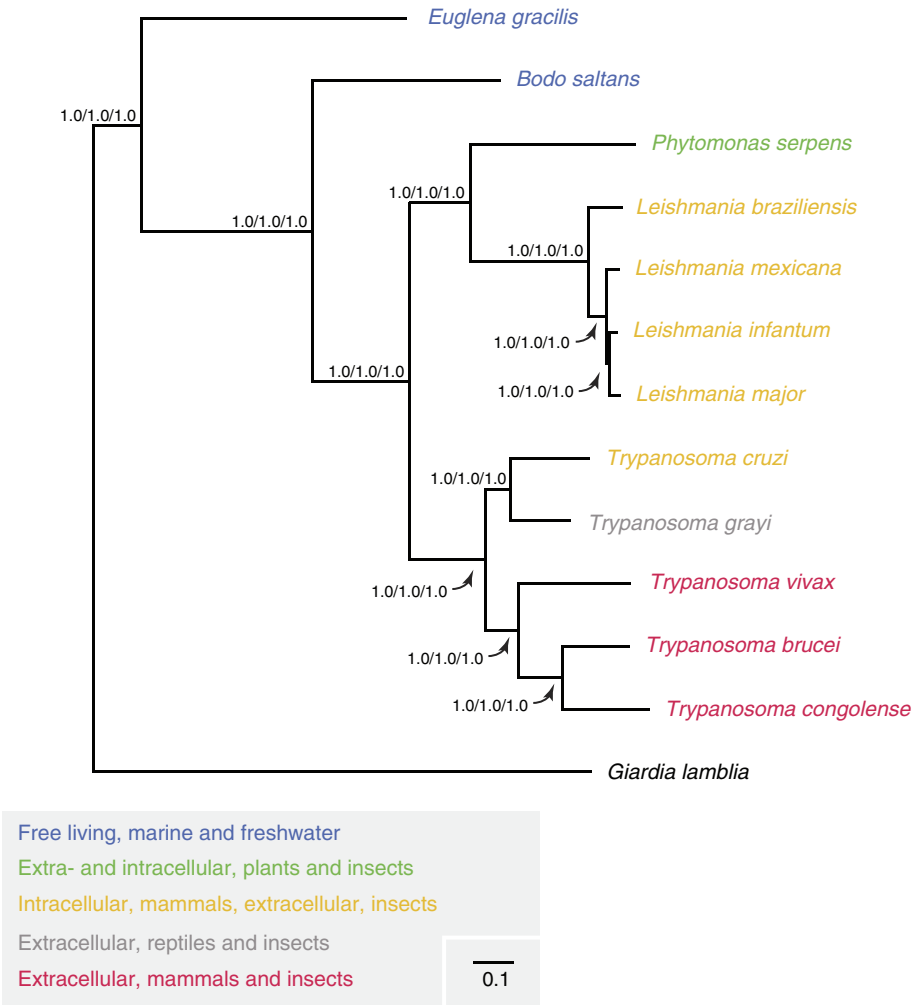

Figure 1. Phylogenetic tree of selected Euglenozoa. Phylogeny is inferred from 119,006 aligned amino acid positions (1,547,078 amino acids) from 959 nuclear genes. The topology was calculated using Bayesian inference (BI), bootstrapped maximum likelihood (ML) analysis and bootstrapped neighbor joining (NJ). ML used the JTT model of amino acid substitution and CAT rates, BI used the WAG model of amino acid substitution and gamma distributed rates approximated by four discrete gamma categories. Branch lengths shown are from the ML topology, scale bar indicates number of changes per site. Values shown at internal nodes represent bootstrap support values for ML and NJ tree as well as posterior probabilities for BI tree. In all cases support for each bipartition in the topology is 100%. *Giardia lamblia*, a diplomonad excavate, is included as an outgroup as the Euglenozoa are also members of the Excavata.

paired-end reads were mapped to the contig set using BWA-MEM[26], paired-end reads that did not map to the assembly were isolated and the above assembly, scaffolding and correction process was repeated until all no-further reads could be assembled. The final draft assembly contained 2,963 sequences greater than 100 bp in length with an N50 of 16.7 kb and a total assembly length of 20.9 Mb and average coverage per assembled contig of ~105X (Figure 2a,b).

ORF finding and annotation

The assembled draft genome of *T. grayi* was subject to gene model prediction using Augustus[27]. In brief, an initial set of gene models were predicted using gene prediction parameters inferred by training Augustus using the set of genes currently annotated in the *T. cruzi* genome. These gene model parameters were used to predict a training set of genes in the draft assembly of *T. grayi*. The training set of genes were then used for multiple iterations of prediction and training until prediction converged on a final set of gene models and no further genes could be detected. The identity of the *T. grayi* DNA used for sequencing was confirmed against database sequences for 18S ribosomal RNA and glycosomal GAPDH genes (AJ005278 and AJ620257 respectively).

Number of predicted genes	Species
3,709	*Trypanosoma cruzi* strain CL Brener
2,109	*Trypanosoma cruzi* marinkellei
1,113	*Trypanosoma cruzi*
694	*Trypanosoma cruzi* Dm28c
194	*Trypanosoma vivax* Y486
164	*Trypanosoma brucei brucei* strain 927/4 GUTat10.1
156	*Trypanosoma brucei gambiense* DAL972
132	*Trypanosoma congolense* IL3000
44	*Angomonas deanei*
36	*Trypanosoma rangeli*
27	*Strigomonas culicis*
26	*Trypanosoma brucei* TREU927
22	*Leishmania major* strain Friedlin
17	*Leishmania braziliensis* MHOM/BR/75/M2904
17	*Leishmania infantum* JPCM5
16	*Leishmania guyanensis*

Table 1. BLASTp similarity scores for *T. grayi* predicted proteins with a bitscore value of >75 using the full non-redundant database from NCBI. Number of 'top hits' with a bitscore value of >75 for each trypanosomatid species are reported. Database was interrogated on 28 February 2014, using BLAST 2.2.27+.

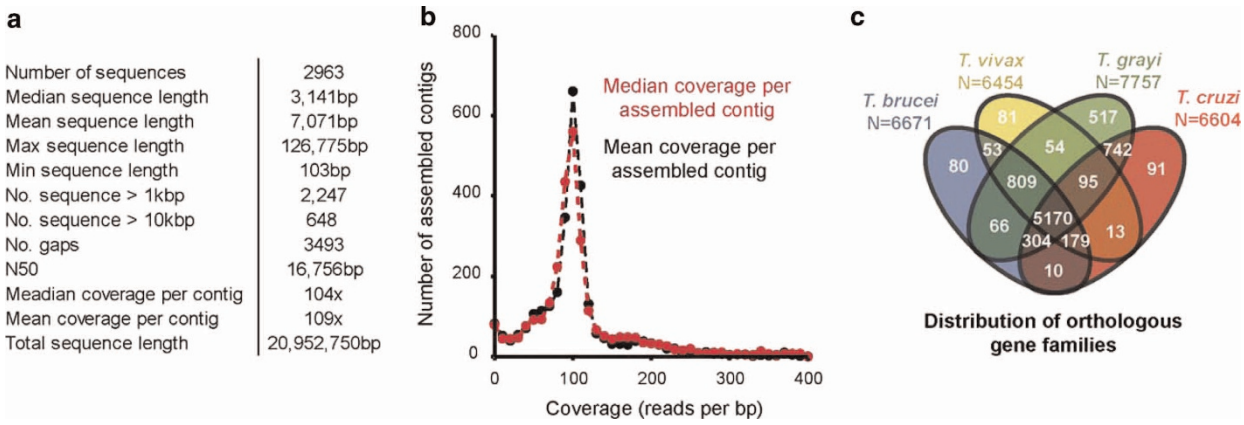

Figure 2. Assembly statistics. (**a**) General assembly statistics describing assembled contig length and coverage. (**b**) Graph showing distribution of coverage estimates for each assembled contig estimated using median and mean coverage depth. (**c**) Venn diagram showing the distribution of orthologous gene families in four of the species used for OrthoMCL clustering.

Gene family analysis

The protein sequence files for a subset of available trypanosomatid genomes were downloaded from TriTrypDB. These were combined with the newly predicted protein sequences from *T. grayi* and subject to orthologue group clustering using OrthoMCL[28]. The presence of gene families in each species was analyzed and the overlap in gene family content between each species and that of the newly assembled

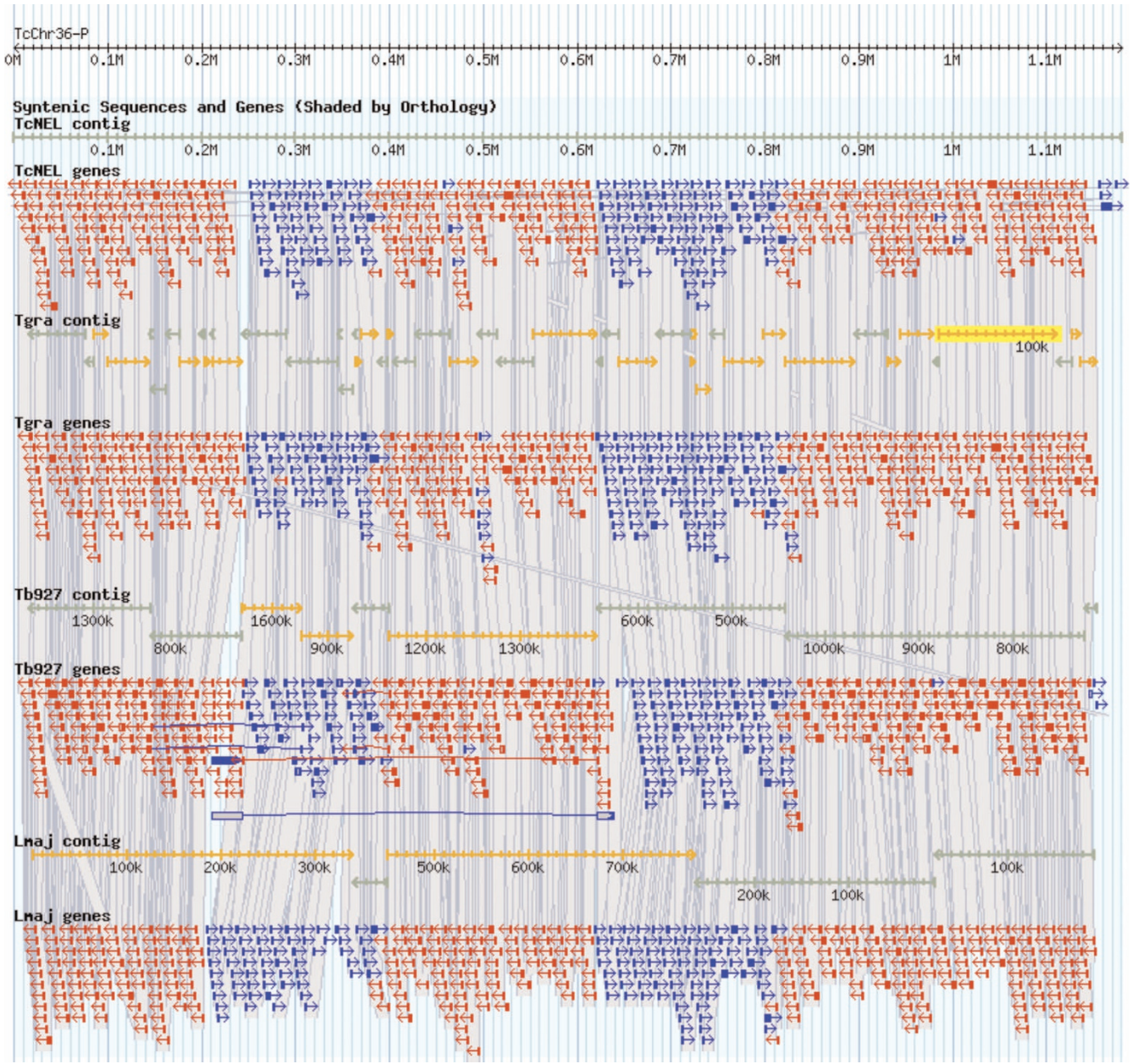

Figure 3. Alignment of ***T. grayi*** contigs against syntenic region of ***T. cruzi*** (Chr 36), ***T. brucei*** and ***L. major***. *T. grayi* assembled contigs were mapped to a contig from Chr 36 of *T. cruzi* (TcNEL, top) in Artemis (http://www.sanger.ac.uk/resources/software/artemis/) together with the equivalent syntenic region from *L. major*. Transcripts (red, blue) are shown beneath mapped contigs (teal, orange) and orthologous sequences are shown as lines behind the main graphic. Despite the fragmentary nature of the *T. grayi* dataset, the data map well to this highly conserved region of the trypanosomatid genomes. Tb927 is the *Trypanosoma brucei* TREU927 genome strain. TcNEL is the *Trypanosoma cruzi* CL Brenner non-Esmeraldo-like genome strain.

T. grayi genome was compared. On average the predicted gene model set of *T. grayi* contained 95% of the gene families present in *T. brucei, T. vivax* and *T. cruzi* (Figure 2c). To put this in context, *T. cruzi* and *T. vivax* contain 84 and 93% of the gene families present in *T. brucei* respectively (Figure 2c).

Phylogenetics for strain verification

Orthologous sequence groups that contained only single copy genes in each of the species that were subject to clustering were selected ($n = 959$). These single copy gene families were aligned using MergeAlign[29] and concatenated to form a super-alignment containing 119,006 aligned amino acid positions across all species (1,547,078 amino acids). This concatenated alignment was subject to

phylogenetic inference using bootstrapped maximum likelihood, Bayesian inference and bootstrapped neighbor joining methods. Maximum likelihood trees were inferred using FastTree[30], utilizing the JTT model of amino acid substitution and CAT rates. A Bayesian inference tree was inferred using MrBayes v3.1.2[31] using the WAG model of amino acid substitution and gamma distributed rates approximated by four discrete gamma categories. Two runs each of four chains were initiated and allowed to run for 200,000 generations sampling every 500 generations. Convergence was assessed through visual inspection of log-likelihood traces and through analysis of the standard deviation of split frequencies. The analysis had reached stationary phase after 15,000 generations and these first 15,000 generations were discarded as burnin prior to inferring the consensus tree. The neighbor joining tree was inferred using QuickTree[32] using the default parameters. The final topology is shown in Fig. 1 and received 100% support at each bipartition from all methods.

Data Records

Data are available both via GenBank as (accession numbers JMRU01000001 to JMRU01002871) and as contigs (accession JMRU00000000.1) under BioProject PRJNA244495, BioSample SAMN02726834 (Data Citation 1). Raw read files are at NCBI SRA under experiment accession SRX620256 and run accession SRR1448313 (Data Citation 2).

Data are also available at TriTrypDB[33] as a hosted genome integrated with other trypanosomatid datasets, http://tritrypdb.org/tritrypdb/showApplication.do (search for all annotated genes), http://tritrypdb.org/common/downloads/Current_Release/TgrayiANR4/ (file download) and http://tritrypdb.org/tritrypdb/getDataset.do?datasets=tgraANR4_primary_genome_RSRC for dataset description.

Technical Validation

The contig statistics of the assembly are reported in Figure 2, and an example region of an assembly against several related trypanosomatid genomes is shown in Figure 3. Phylogenetic strain validation as described above confirmed the placement of *T. grayi* ANR4 with other species of genus *Trypanosoma* (Fig. 2) and identity of the sequenced genome here with the previously reported 18S and glycosomal GAPDH genes (AJ005278 and AJ620257 respectively). The phylogenomic position of *T. grayi* closer to *T. cruzi* than *T. brucei* is also supported by BLASTp analysis of all predicted open reading frames (Table 1).

References

1. Simpson, A. G., Stevens, J. R. & Lukes, J. The evolution and diversity of kinetoplastid flagellates. *Trends Parasitol.* **22,** 168–174 (2006).
2. Flegontov, P. *et al.* Paratrypanosoma is a novel early-branching trypanosomatid. *Curr. Biol.* **23,** 1787–1793 (2013).
3. Nagajyothi, F. *et al.* Mechanisms of *Trypanosoma cruzi* persistence in Chagas disease. *Cell. Micro.* **14,** 634–643 (2012).
4. Denkers, E. Y. & Butcher, B. A. Sabotage and exploitation in macrophages parasitized by intracellular protozoans. *Trends Parasitol.* **21,** 35–41 (2005).
5. Rudenko, G. African trypanosomes: the genome and adaptations for immune evasion. *Essays Biochem.* **51,** 47–62 (2011).
6. Berriman, M. *et al.* The genome of the African trypanosome *Trypanosoma brucei*. *Science* **309,** 416–422 (2005).
7. Ivens, A. C. *et al.* The genome of the kinetoplastid parasite, *Leishmania major*. *Science* **309,** 436–442 (2005).
8. El-Sayed, N. M. *et al.* The genome sequence of *Trypanosoma cruzi,* etiologic agent of Chagas disease. *Science* **309,** 409–415 (2005).
9. Peacock, C. S. *et al.* Comparative genomic analysis of three Leishmania species that cause diverse human disease. *Nat. Genet.* **39,** 839–847 (2007).
10. Raymond, F. *et al.* Genome sequencing of the lizard parasite *Leishmania tarentolae* reveals loss of genes associated to the intracellular stage of human pathogenic species. *Nucleic Acids Res.* **40,** 1131–1147 (2012).
11. Downing, T. *et al.* Whole genome sequencing of multiple *Leishmania donovani* clinical isolates provides insights into population structure and mechanisms of drug resistance. *Genome Res.* **21,** 2143–2156 (2011).
12. Jackson, A. P. *et al.* Antigenic diversity is generated by distinct evolutionary mechanisms in African trypanosome species. *Proc. Natl Acad. Sci. USA* **109,** 3416–3421 (2012).
13. Jackson, A. P. *et al.* A cell-surface phylome for African trypanosomes. *PLoS NTD* **7,** e2121 (2013).
14. Franzén, O. *et al.* Shotgun sequencing analysis of *Trypanosoma cruzi* I Sylvio X10/1 and comparison with *T. cruzi* VI CL Brener. *PLoS NTD* **5,** e984 (2011).
15. Ackermann, A. A., Panunzi, L. G., Cosentino, R. O., Sánchez, D. O. & Agüero, F. A genomic scale map of genetic diversity in *Trypanosoma cruzi*. *BMC Genom.* **13,** 736 (2012).
16. Fermino, B. R. *et al.* The phylogeography of trypanosomes from South American alligatorids and African crocodilids is consistent with the geological history of South American river basins and the transoceanic dispersal of *Crocodylus* at the Miocene. *Parasites Vectors* **6,** 313 (2013).
17. Hoare, C. A. Studies on *Trypanosoma grayi* II. Experimental transmission to the crocodile. *Trans. Roy. Soc. Trop. Med. Hyg.* **23,** 39–56 (1929).
18. Hoare, C. A. Studies on *Trypanosoma grayi*. III. Life-Cycle in the Tsetse-fly and in the Crocodile. *Parasitology* **23,** 449 (1929).
19. Manna, P. T., Kelly, S. & Field, M. C. Adaptin evolution in kinetoplastids and emergence of the variant surface glycoprotein coat in African trypanosomatids. *Mol. Phylogen. Evol.* **67,** 123–128 (2013).
20. Stevens, J. R., Noyes, H. A., Dover, G. A. & Gibson, W. C. The ancient and divergent origins of the human pathogenic trypanosomes, *Trypanosoma brucei* and *T. cruzi*. *Parasitology* **118,** 107–116 (1999).
21. Hamilton, P. B., Gibson, W. C. & Stevens, J. R. Patterns of co-evolution between trypanosomes and their hosts deduced from ribosomal RNA and protein-coding gene phylogenies. *Mol. Phylogen. Evol.* **44,** 15–25 (2007).
22. Hamilton, P. B., Stevens, J. R., Gaunt, M. W., Gidley, J. & Gibson, W. C. Trypanosomes are monophyletic: evidence from genes for glyceraldehyde phosphate dehydrogenase and small subunit ribosomal RNA. *Int. J. Parasitol.* **34,** 1393–1404 (2004).
23. Lohse, M. *et al.* RobiNA: a user-friendly, integrated software solution for RNA-Seq-based transcriptomics. *Nucleic Acids Res.* **40,** W622–W627 (2012).

24. Gnerre, S. *et al.* High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proc. Natl Acad. Sci. USA* **108,** 1513–1518 (2011).
25. Simpson, J. T. & Durbin, R. Efficient construction of an assembly string graph using the FM-index. *Bioinformatics* **26,** i367–i373 (2010).
26. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler Transform. *Bioinformatics* **26,** 589–595 (2010).
27. Keller, O., Kollmar, M., Stanke, M. & Waack, S. A novel hybrid gene prediction method employing protein multiple sequence alignments. *Bioinformatics* **27,** 757–763 (2011).
28. Li, L., Stoeckert, C. J. & Roos, D. S. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.* **13,** 2178–2189 (2003).
29. Collingridge, P. W. & Kelly, S. MergeAlign: improving multiple sequence alignment performance by dynamic reconstruction of consensus multiple sequence alignments. *BMC Bioinformatics.* **13,** 117 (2012).
30. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2—approximately maximum-likelihood trees for large alignments. *PLoS ONE* **5,** e9490 (2010).
31. Ronquist, F. & Huelsenbeck, J. P. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19,** 1572–1574 (2003).
32. Howe, K., Bateman, A. & Durbin, R. QuickTree: building huge Neighbour-Joining trees of protein sequences. *Bioinformatics* **18,** 1546–1547 (2002).
33. Aslett, M. *et al.* TriTrypDB: a functional genomic resource for the Trypanosomatidae. *Nucleic Acids Res.* **38,** D457–D462 (2010).

Data Citations

1. Kelly, S., Ivens, A., Manna, P. T., Gibson, W. & Field, M. C. GenBank PRJNA244495 (2014).
2. Kelly, S., Ivens, A., Manna, P. T., Gibson, W. & Field, M. C. NCBI Sequence Read Archive SRX620256 (2014).

Acknowledgements

We thank Omar Harb (Philadelphia) for integration of data into TriTrypDB and also for generating Figure 3. This work was supported in part by the Wellcome Trust (program grant 082813 to MCF). S.K. is a Leverhulme Trust early career Fellow. This Whole Genome Shotgun project has been deposited at DDBJ/EMBL/GenBank under the accession JMRU00000000. The version described in this paper is version JMRU01000000. M.C.F. had full access to all the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

Author Contributions

S.K., created assemblies, predicted the gene models and built the phylogeny, edited the manuscript, A.I., ORF annotations and data processing for NCBI submission, edited the manuscript, P.M., coordinated the project, isolated DNA, wrote the manuscript, W.G., provided DNA, edited the manuscript, M.C.F., conceived/coordinated the project, edited the manuscript.

Additional information

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Kelly, S. *et al.* A draft genome for the African crocodilian trypanosome *Trypanosoma grayi*. *Sci. Data* 1:140024 doi: 10.1038/sdata.2014.24 (2014).