July 29, 2021 revised

Genome assemblies across the diverse evolutionary spectrum of *Leishmania* protozoan parasites

Wesley C. Warren[1], Natalia S. Akopyants[2], Deborah E. Dobson[2], Christiane Hertz-Fowler[3], Lon-Fye Lye[2], Peter J. Myler[4,5], Gowthaman Ramasamy[4], Achchuthan Shanmugasundram[3], Fatima Silva-Franco[3], Sascha Steinbiss[6], Chad Tomlinson[7], Richard K. Wilson[8], and Stephen M. Beverley[2,*]

[1] Department of Animal Sciences, Institute of Data Science and Informatics, Bond Life Sciences Center, University of Missouri, Columbia, MO, USA.

[2] Department of Molecular Microbiology, Washington University School of Medicine, St Louis, MO; Washington University School of Medicine, St Louis, MO.

[3] Institute of Integrative Biology, Biosciences Building, University of Liverpool, Crown Street, Liverpool L69 7ZB, UK.

[4] Center for Global Infectious Disease Research, Seattle Children's Research Institute, Seattle, WA, USA.

[5] Department of Pediatrics, Department of Biomedical Informatics and Medical Education and Department of Global Health, University of Washington, Seattle WA, USA

[6] Wellcome Sanger Institute, Hinxton, Cambridgeshire, UK, CB10 1SA

[7] McDonnell Genome Institute, Washington University School of Medicine, St Louis, MO; USA

[8] The Steve and Cindy Rasmussen Institute for Genomic Medicine, Nationwide Children's Hospital, Columbus, Ohio, USA.

*Corresponding author: Stephen M. Beverley, Molecular Microbiology, Washington University School of Medicine, St Louis, MO 63110, USA. Telephone 1-314-747-2630; email stephen.beverley@wustl.edu.

Keywords: *Leishmania*, *Viannia*, *Mundinia*, comparative genomics, evolution of parasitism, Kinetoplastida, Trypanosomatidae.

Running title: Genome assemblies across *Leishmania* phylogeny

Abstract

We report high-quality draft assemblies and gene annotation for 13 species and/or strains of the protozoan parasite genera *Leishmania, Endotrypanum* and *Crithidia*, which span the phylogenetic diversity of the Leishmaniinae sub-family within the kinetoplastid order of the phylum Euglenazoa. These resources will support studies on the origins of parasitism.

Leishmania species are widespread parasites of mammals transmitted by biting insects. Over 1.7 billion people worldwide are at risk, with hundreds of millions of people infected (1-4). The genus comprises more than 50 species, which are primarily zoonotic but in humans cause disease ranging from mild cutaneous lesions to more disseminated forms, to fatal visceralizing disease (5). While parasitism by *Leishmania* has been intensively studied, the species-specific factors that enable mammalian or insect host infections are less well understood. To provide a broad phylogenetic snapshot, we selected a spectrum of species and strains across the Leishmaniinae sub-family (5), targeting lineages within the subgenera *Leishmania*, *Viannia*, and *Mundinia* , as well as allied *Endotrypanum* and the outgroup *Crithidia fasciculata* (Table 1).

Parasites were cultivated in M199 or Schneider's medium (6) and grown to late log phase, before harvesting, lysis, and DNA purification by phenol/chloroform extraction and/or banding in CsCl gradients (to remove mitochondrial maxi- or minicircle DNA). Sequencing libraries were generated using the Illumina paired-end DNA sample preparation kit (PE-102-1001) according to the manufacturer's directions. Fragment libraries of 3 and 8kb were prepared using protocols for 454 sequencing (Roche Life Sciences). Sequencing was performed on either a 454 GS FLX Titanium (average read length of 305bp; Roche 454 Life Sciences), or Illumina GAIIx and HiSeq2000 instruments (paired end 100bp read length), except for *Crithidia* which additionally utilized long-reads generated on a RSII instrument (P5/C3 chemistry; Pacific BioSciences) (7). Total sequence genome coverage on the Illumina GAIIX or HiSeq 2000 instrument was on average 105x with tiered library insert sizes (50x fragments, 45x 3kb, and 10x 8kb and 0.05x 40kb). For all Illumina sequences we used the read processing steps within the ALLPATHS-LG (8) software prior to *de novo* assembly, which incorporates read error correction methods described by Pevzner *et al* (9). Genome assemblies were performed with default parameters using Newbler v2.0.1 (10) for 454 reads, ALLPATHS-LG (8) for Illumina reads, or HGAP version 3 (11) for long-reads (Table 1). Contigs and scaffolds were organized into pseudochromosomes using ABACAS2 (https://github.com/satta/ABACAS2), a successor to ABACAS (12), by alignment to the *Leishmania major* Friedlin genome, with the exception of *L.*

braziliensis M2903 and *L. panamensis*, which were aligned to the *L. braziliensis* M2904 genome. The estimated haploid genome sizes ranged from 30.4 to 41.3 megabases (13).

Gene annotations were performed using the comprehensive Companion tool which incorporates a variety of *de novo* prediction criteria, as well as information from closely related genomes when available (14). The number of protein-coding genes predicted ranged from 8,285 to 9,619, typical of other *Leishmania* (13). Full annotations, as well as a variety of tools for the visualization or analysis of these genomes, are available from TriTrypDB (www.tritrypdb.org).

Public availability and accession number(s). Assemblies are deposited in the NCBI GenBank Repository under the Bioproject numbers in Table 1, including links to the primary data and annotations (PRJNA50303, PRJNA50301, PRJNA192717, PRJNA192712, PRJNA192710, PRJNA169676, PRJNA169673, PRJNA165955, PRJNA165959, PRJNA192711, PRJNA192703, PRJNA165953, PRJNA165885). Chromosome builds are available through the TriTrypDB portal (http://tritrypdb.org/tritrypdb)**.**

Acknowledgements.

NIH grant AI29646 to Stephen M. Beverley provided support for NSA, DED and L-FL, NIH-NHGRI grant HG00307907 to Richard K. Wilson supported WCW, and CT, and AI103858 to Peter J Myler supported GA. Funding to Christiane Hertz-Fowler from Wellcome Trust grants WT099198MA and WT108443MA provided support for FS-F, AS and SS. We thank Patrick Minx for assembly curation. We thank the colleagues listed in Table 1 for generously providing strains and Brian Brunk, David Roos, Thomas Otto and Matt Berriman for discussions. The WU Institutional Biosafety Committee reviewed and approved the parasite work reported here (01-015).

References.

1. Pigott DM, Bhatt S, Golding N, Duda KA, Battle KE, Brady OJ, Messina JP, Balard Y, Bastien P, Pratlong F, Brownstein JS, Freifeld CC, Mekaru SR, Gething PW, George DB, Myers MF, Reithinger R, Hay SI. 2014. Global distribution maps of the leishmaniases. Elife 3:e28051.

2. Alvar J, Velez ID, Bern C, Herrero M, Desjeux P, Cano J, Jannin J, den Boer M. 2012. Leishmaniasis worldwide and global estimates of its incidence. PLoS One 7:e35671.

3. Banuls AL, Bastien P, Pomares C, Arevalo J, Fisa R, Hide M. 2011. Clinical pleiomorphism in human leishmaniases, with special mention of asymptomatic infection. Clin Microbiol Infect 17:1451-61.

4. Singh OP, Hasker E, Sacks D, Boelaert M, Sundar S. 2014. Asymptomatic Leishmania infection: a new challenge for *Leishmania* control. Clin Infect Dis 58:1424-9.

5. Espinosa OA, Serrano MG, Camargo EP, Teixeira MMG, Shaw JJ. 2018. An appraisal of the taxonomy and nomenclature of trypanosomatids presently classified as *Leishmania* and *Endotrypanum*. Parasitology 145:430-442.

6. Lye LF, Owens K, Shi H, Murta SM, Vieira AC, Turco SJ, Tschudi C, Ullu E, Beverley SM. 2010. Retention and loss of RNA interference pathways in trypanosomatid protozoans. PLoS Pathog 6:e1001161.

7. Filosa JN, Berry CT, Ruthel G, Beverley SM, Warren WC, Tomlinson C, Myler PJ, Dudkin EA, Povelones ML, Povelones M. 2019. Dramatic changes in gene expression in different forms of *Crithidia fasciculata* reveal potential mechanisms for insect-specific adhesion in kinetoplastid parasites. PLoS Negl Trop Dis 13:e0007570.

8. Gnerre S, Maccallum I, Przybylski D, Ribeiro FJ, Burton JN, Walker BJ, Sharpe T, Hall G, Shea TP, Sykes S, Berlin AM, Aird D, Costello M, Daza R, Williams L, Nicol R, Gnirke A, Nusbaum C, Lander ES, Jaffe DB. 2011. High-quality draft

assemblies of mammalian genomes from massively parallel sequence data. Proc Natl Acad Sci U S A 108:1513-8.

9. Pevzner PA, Tang H, Waterman MS. 2001. An Eulerian path approach to DNA fragment assembly. Proc Natl Acad Sci U S A 98:9748-53.

10. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z, Dewell SB, Du L, Fierro JM, Gomes XV, Godwin BC, He W, Helgesen S, Ho CH, Irzyk GP, Jando SC, Alenquer ML, Jarvie TP, Jirage KB, Kim JB, Knight JR, Lanza JR, Leamon JH, Lefkowitz SM, Lei M, Li J, Lohman KL, Lu H, Makhijani VB, McDade KE, McKenna MP, Myers EW, Nickerson E, Nobile JR, Plant R, Puc BP, Ronan MT, Roth GT, Sarkis GJ, Simons JF, Simpson JW, Srinivasan M, Tartaro KR, Tomasz A, Vogt KA, Volkmer GA, et al. 2005. Genome sequencing in microfabricated high-density picolitre reactors. Nature 437:376-80.

11. Chaisson MJ, Huddleston J, Dennis MY, Sudmant PH, Malig M, Hormozdiari F, Antonacci F, Surti U, Sandstrom R, Boitano M, Landolin JM, Stamatoyannopoulos JA, Hunkapiller MW, Korlach J, Eichler EE. 2015. Resolving the complexity of the human genome using single-molecule sequencing. Nature 517:608-11.

12. Assefa S, Keane TM, Otto TD, Newbold C, Berriman M. 2009. ABACAS: algorithm-based automatic contiguation of assembled sequences. Bioinformatics 25:1968-9.

13. Ivens AC, Peacock CS, Worthey EA, Murphy L, Aggarwal G, Berriman M, Sisk E, Rajandream MA, Adlem E, Aert R, Anupama A, Apostolou Z, Attipoe P, Bason N, Bauser C, Beck A, Beverley SM, Bianchettin G, Borzym K, Bothe G, Bruschi CV, Collins M, Cadag E, Ciarloni L, Clayton C, Coulson RM, Cronin A, Cruz AK, Davies RM, De Gaudenzi J, Dobson DE, Duesterhoeft A, Fazelina G, Fosker N, Frasch AC, Fraser A, Fuchs M, Gabel C, Goble A, Goffeau A, Harris D, Hertz-Fowler C, Hilbert H, Horn D, Huang Y, Klages S, Knights A, Kube M, Larke N, Litvin L, et al. 2005. The genome of the kinetoplastid parasite, *Leishmania major*. Science 309:436-42.

14. Steinbiss S, Silva-Franco F, Brunk B, Foth B, Hertz-Fowler C, Berriman M, Otto TD. 2016. Companion: a web server for annotation and analysis of parasite genomes. Nucleic Acids Res 44:W29-34.

Table 1. Description of *Leishmania* species and strains including assembly parameters and links.

Subgenus	Species	strain ID	WHO Code	Source	Provenance	NCBI GenBank Bioproject[1]	Sequencing Platforms	Assembler	Contigs	N50 contig length	Assembly size (bp)	%GC	Pseudo-chromo-somes	Protein coding genes	Other assemblies
L. Leishmania	*L. major*	SD75.1 (clone)	MHOM/SN/74/SD	human, cutaneous	D. Sacks, Bethesda USA	PRJNA50303	454, Illumina	Newbler	891	95,380	31,727,271	59.3	36	8,818	
		LV39 (clone5)	MHOM/Sv/59/P	human, cutaneous	R. Titus, Boston, USA	PRJNA50301	454, Illumina	Newbler	1,667	71,452	31,961,985	59.5	36	8,971	
	L. gerbilli	LEM452	MRHO/CN/60/GERBILLI	gerbil	P. Bastien, J-P Dedet, Montpellier, France	PRJNA192717	454, Illumina	Allpaths	1,248	57,008	30,822,621	59.6	36	8,599	
	L. turanica	LEM563	MMEL/SU/79/MEL	gerbil	P. Bastien, J-P Dedet, Montpellier, France	PRJNA192712	454, Illumina	Allpaths	1,669	39,210	30,876,294	59.5	36	8,608	
	L. arabica	LEM1108	MPSA/SA/83/JISH220	*Psammomys*	P. Bastien, J-P Dedet, Montpellier, France	PRJNA192710	454, Illumina	Allpaths	1,530	52,119	30,774,332	59.2	36	8,646	
	L. tropica	L590	MHOM/IL/1990/P283	human, cutaneous	C. Jaffe, Jerusalem, Israel	PRJNA169676	454, Illumina	Allpaths	1,938	32,739	31,326,083	59.6	36	8,824	(12)
	L. aethiopica	L147	MHOM/ET/1972/L100	human, diffuse cutaneous, relapsing	C. Jaffe, Jerusalem, Israel	PRJNA169673	454, Illumina	Allpaths	1,758	38,498	31,026,739	60.1	36	8,722	
L. Viannia	*L.braziliensis*	M2903	MHOM/BR/72/M2903	human, cutaneous	J. Shaw, Brasil	PRJNA165955	454, Illumina	Newbler	3,934	61,918	32,590,753	57.4	35	9,269	
	L. panamensis	L13	MHOM/COL/81/L13	human, mucosal	N. Saravia, Cali, Colombia	PRJNA165959	454, Illumina	Allpaths	3,163	22,576	31,108,242	57.4	35	8,665	
L. Mundinia	*L. enriettii*	LEM3045	MCAV/BR/95/CUR3	*Cavia porcellus* guinea pig	P. Bastien, J-P Dedet, Montpellier, France	PRJNA192711	454, Illumina	Allpaths	1,171	102,666	30,427,298	59.3	36	8,731	(14)
	L. martiniquensis	LEM2494	MHOM/MQ/92/MAR1	human, diffuse cutaneous, HIV	P. Bastien, J-P Dedet, Montpellier, France	PRJNA192703	454, Illumina	Allpaths	628	147,290	30,528,357	59.6	36	8,483	(14)
Endotrypanum	*Endotrypanum monterogei*	LV88	none	*Choloepus hoffmani* sloth	Michael Chance, Liverpool UK	PRJNA165953	454, Illumina	Newbler	3,517	33,059	32,086,870	52.5	36	8,285	(13)
Crithidia	*Crithidia fasciculata*	Cf-C1 (clone)	none	mosquito	Larry Simpson, Los Angeles, USA	PRJNA165885	454, Illumina, PacBio	HGAP	494	778,443	41,297,378	57.0	30	9,619	(8)

Footnote 1 Each bioproject link contains links to the current assembly, primary data sets, and other relevant information.