



OPEN

DATA DESCRIPTOR

Chromosome-scale genome sequencing, assembly and annotation of six genomes from subfamily *Leishmaniinae*

Hatim Almutairi^{1,2}, Michael D. Urbaniak¹, Michelle D. Bates¹, Narissara Jariyapan³, Godwin Kwakye-Nuako⁴, Vanete Thomaz Soccol⁵, Waleed S. Al-Salem², Rod J. Dillon¹, Paul A. Bates¹ & Derek Gatherer¹✉

We provide the raw and processed data produced during the genome sequencing of isolates from six species of parasites from the sub-family *Leishmaniinae*: *Leishmania martiniquensis* (Thailand), *Leishmania orientalis* (Thailand), *Leishmania enriettii* (Brazil), *Leishmania* sp. Ghana, *Leishmania* sp. Namibia and *Porcisia hertigi* (Panama). *De novo* assembly was performed using Nanopore long reads to construct chromosome backbone scaffolds. We then corrected erroneous base calling by mapping short Illumina paired-end reads onto the initial assembly. Data has been deposited at NCBI as follows: raw sequencing output in the Sequence Read Archive, finished genomes in GenBank, and ancillary data in BioSample and BioProject. Derived data such as quality scoring, SAM files, genome annotations and repeat sequence lists have been deposited in Lancaster University's electronic data archive with DOIs provided for each item. Our coding workflow has been deposited in GitHub and Zenodo repositories. This data constitutes a resource for the comparative genomics of parasites and for further applications in general and clinical parasitology.

Background & Summary

Leishmaniasis is a neglected tropical disease. It is considered to be a disease of poverty, primarily affecting low and middle-income countries (LMICs). Leishmaniasis is caused by parasites of the genus *Leishmania* and 18 different species are known to infect humans¹. 98 sandfly species are suspected or confirmed vectors of *Leishmania*². There are three major types of leishmaniasis: visceral, also known as kala-azar, is fatal if left untreated in over 95% of cases; cutaneous, the most common form, causes skin lesions leaving life-long scars and serious disability or stigma; mucocutaneous, leads to partial or total destruction of mucous membranes of the nose, mouth and throat³. Over one billion people live in endemic areas and are at risk of leishmaniasis. It is estimated that each year, globally, new cases of cutaneous leishmaniasis occur at an incidence of 700,000 to 1.2 million or more in over 100 countries⁴. Additionally, up to 300,000 visceral leishmaniasis cases cause more than 200,000 deaths annually⁵.

The genus *Leishmania* is divided into four subgenera: *L. Leishmania*, *L. Viannia*, *L. Sauroleishmania* and the newest subgenus *L. Mundinia*, the latter now accommodating several species from the *L. enriettii* complex and others, from five continents^{6–12}. In 1994, the *Leishmania* Genome Network was initiated¹³ and announced, ten years later, the assembly of the *Leishmania major* Friedlin strain as the first *Leishmania* reference genome¹⁴. Since then, a total of 58 genomes have become available publicly, assembled at a variety of levels of completeness ranging from contigs to chromosome level. Prior to our project, only two *L. Mundinia* subgenus genomes have been sequenced and assembled: *Leishmania enriettii*, strain LEM3045 (GCA_000410755) and *Leishmania* sp. MAR, strain LEM2494 (GCA_000410755). The genus *Porcisia* is a sister genus of *Leishmania* within the

¹Division of Biomedical & Life Sciences, Faculty of Health & Medicine, Lancaster University, Lancaster, LA1 4YT, UK. ²Ministry of Health, Riyadh, Saudi Arabia. ³Department of Parasitology, Faculty of Medicine, Chulalongkorn University, Bangkok, 10330, Thailand. ⁴Department of Biomedical Sciences, School of Allied Health Sciences, College of Health & Allied Sciences, University of Cape Coast, Cape Coast, Ghana. ⁵Laboratório de Biologia Molecular, Programa de Pós Graduação em Engenharia de Bioprocessos e Biotecnologia, Universidade Federal do Paraná, Curitiba, Brazil. ✉e-mail: d.gatherer@lancaster.ac.uk

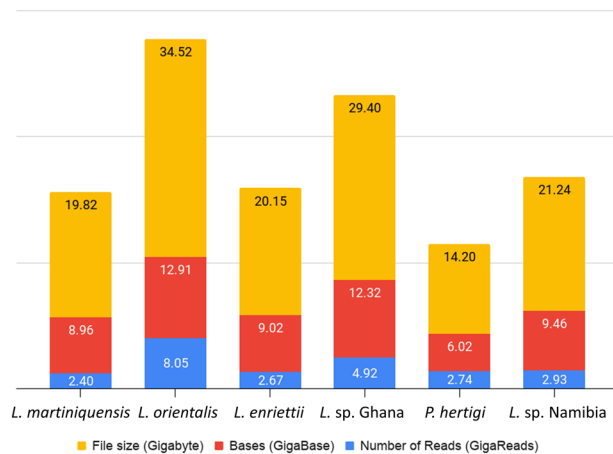


Fig. 1 Stacked column chart showing number of sequenced reads in GigaReads (blue), number of yielded bases in GigaBases (red), and the file sizes in Gigabytes (yellow) for each genome assembly.

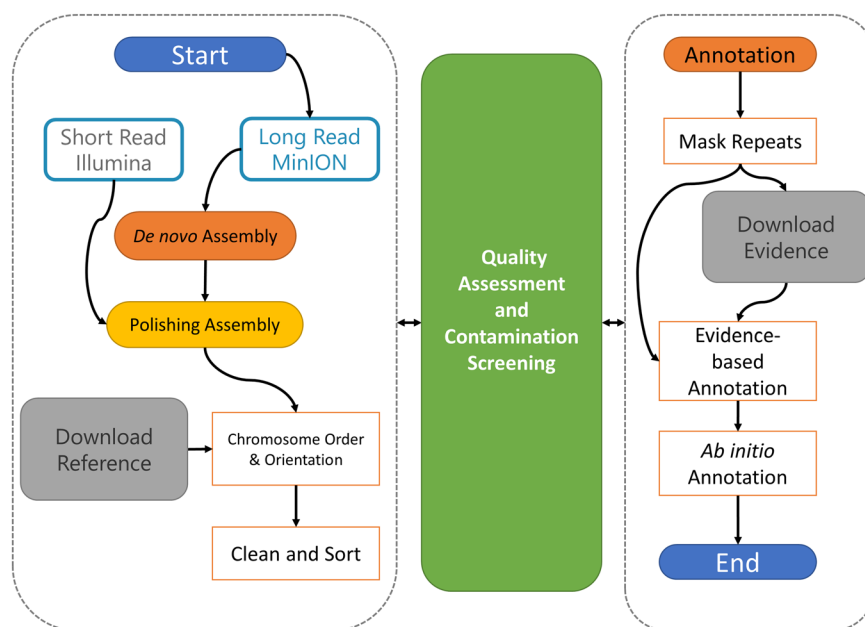


Fig. 2 Flowchart showing the analysis workflow strategy.

Sample	Strain	Isolate	BioSample	BioProject
<i>L. (Mundinia) martiniquensis</i>	LV760	LSCM1	SAMN17294109	PRJNA691531
<i>L. (Mundinia) orientalis</i>	LV768	LSCM4	SAMN17294111	PRJNA691532
<i>L. (Mundinia) enriettii</i>	LV763	CUR178	SAMN17294112	PRJNA691534
<i>L. (Mundinia) sp. Ghana</i>	LV757	GH5	SAMN17294115	PRJNA691536
<i>L. (Mundinia) sp. Namibia</i>	LV425	253	SAMN17294129	PRJNA689706
<i>Porcisia hertigi</i>	LV43	C119	SAMN17294121	PRJNA691541

Table 1. Sample descriptions for all assemblies.

sub-family *Leishmaniinae*. Prior to the release of our genome, there were no genome sequences for genus *Porcisia*. Subsequently, the partial genome of *P. deanei* was released and published¹⁵.

We assembled and annotated the genomes of five *L. Mundinia* species – those of *L. martiniquensis*, *L. orientalis*, *L. enriettii*, *L. sp. Ghana* and *L. sp. Namibia* - and one genome in the genus *Porcisia* – that of *P. hertigi*, formerly known as *L. hertigi*¹⁶ - using Illumina and Nanopore sequencing. The two isolates from Ghana and Namibia are from new species that have not yet been formally named. The World Health Organization (WHO)

Tool	Website	conda or docker link
AGAT	https://github.com/NBISweden/AGAT	https://anaconda.org/conda-forge/agate
AUGUSTUS	http://bioinf.uni-greifswald.de/webaugustus/about	https://hub.docker.com/r/hatimalmutairi/lmgaap-maker
BCFtools	http://samtools.github.io/bcftools/	https://anaconda.org/bioconda/bcftools
bedtools	https://bedtools.readthedocs.io/en/latest/	https://anaconda.org/bioconda/bedtools
blast+	https://blast.ncbi.nlm.nih.gov/Blast.cgi	https://anaconda.org/bioconda/blast
FastQC	https://www.bioinformatics.babraham.ac.uk/projects/fastqc/	https://anaconda.org/bioconda/fastqc
Flye	https://github.com/fenderglass/Flye	https://anaconda.org/bioconda/flye
funannotate	https://github.com/nextgenusfs/funannotate	https://anaconda.org/bioconda/funannotate
GAAS	https://github.com/NBISweden/GAAS	https://anaconda.org/bioconda/gaas
GeneMark	http://exon.gatech.edu/GeneMark/	https://hub.docker.com/r/hatimalmutairi/lmgaap-maker
Genometools	http://genometools.org/	https://anaconda.org/bioconda/genometools-genometools
interproscan	https://www.ebi.ac.uk/interpro/search/sequence/	https://hub.docker.com/r/blaxterlab/interproscan
MAKER2	https://www.yandell-lab.org/software/maker.html	https://hub.docker.com/r/hatimalmutairi/lmgaap-maker
minimap2	https://github.com/lh3/minimap2	https://anaconda.org/bioconda/minimap2
MultiQC	https://multiqc.info/	https://anaconda.org/bioconda/multiqc
MUMmer	http://mummer.sourceforge.net/	https://anaconda.org/bioconda/mummer
Pilon	https://github.com/broadinstitute/pilon/wiki	https://anaconda.org/bioconda/pilon
pycoQC	https://pypi.org/project/pycoQC/	https://anaconda.org/bioconda/pycoqc
RaGOO	https://github.com/malonge/RaGOO	https://anaconda.org/imperial-college-research-computing/ragoo
RepeatMasker	http://www.repeatmasker.org/	https://hub.docker.com/r/hatimalmutairi/lmgaap-maker
SAMtools	https://github.com/samtools/samtools	https://anaconda.org/bioconda/samtools
Snakemake	https://snakemake.readthedocs.io/en/stable/	https://anaconda.org/bioconda/snakemake
TEclass	http://www.compgen.uni-muenster.de/tools/teclass/index.hbi?lang=en	https://hub.docker.com/r/hatimalmutairi/teclass-2.1.3b
wordcloud	Not available	https://anaconda.org/conda-forge/wordcloud

Table 2. Tools used in analysis workflow with conda or docker link.

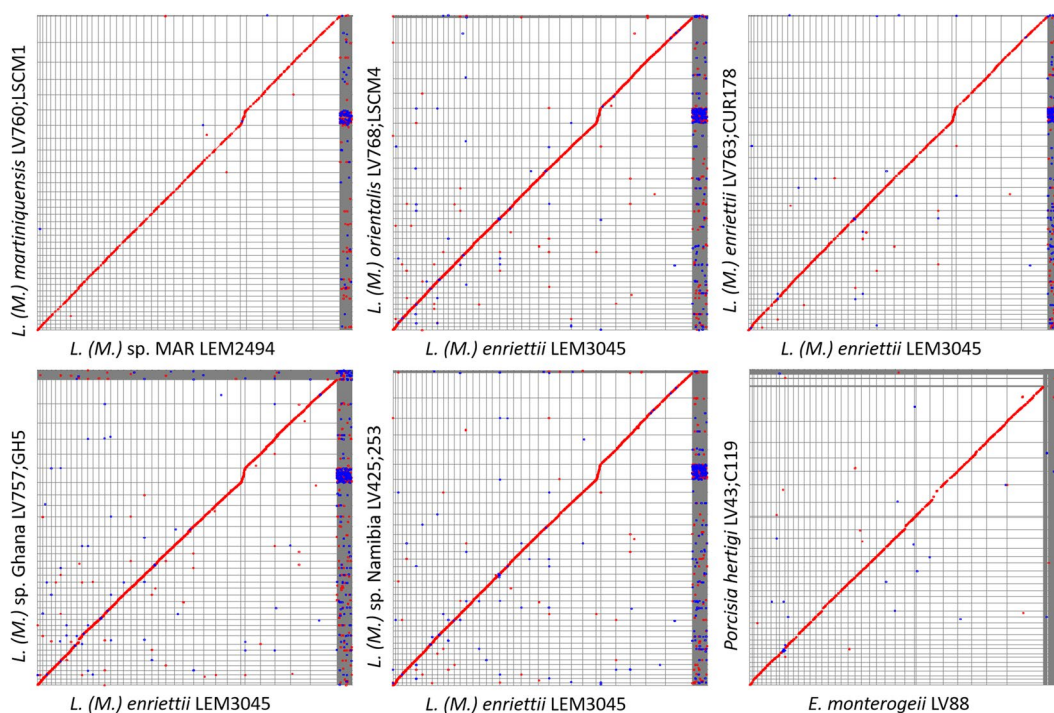


Fig. 3 Dotplot representing synteny between each of our genomes and its wordcloud-predicted closest related reference genome, produced using MUMmer.

codes for the six isolates are: *L. martiniquensis* MHOM/TH/2012/LSCM1;LV760; *L. orientalis* MHOM/TH/2014/LSCM4;LV768; *L. enriettii* MCAV/BR/2001/CUR178;LV673; *L. sp. Ghana* MHOM/GH/2012/GH5;LV757; *L. sp. Namibia* MPRO/NA/1975/252;LV425; and *P. hertigi* MCOE/PA/1965/C119;LV43. Nanopore long reads were

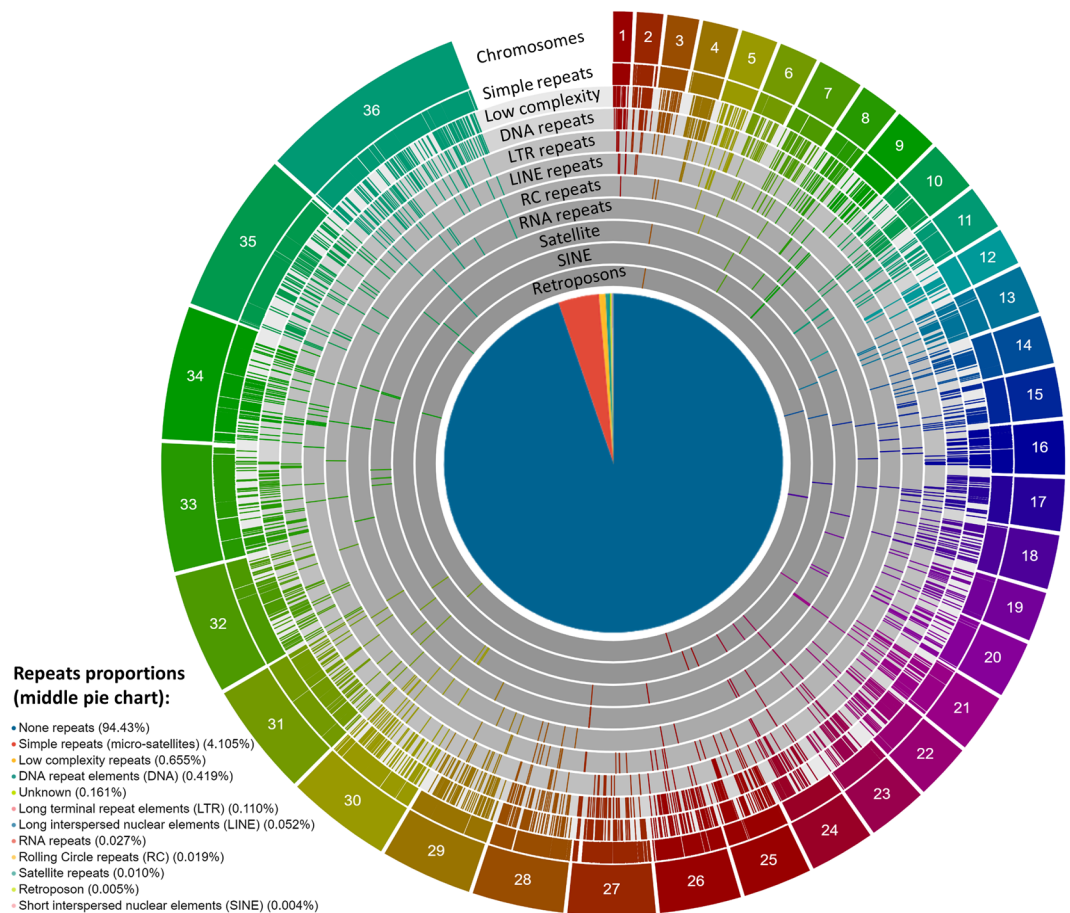


Fig. 4 Example genome-wide repeat plot for *L. martiniquensis*, stratified: simple (micro-satellites), low complexity, DNA, long terminal repeats (LTRs), long interspersed nuclear elements (LINEs), RNA, rolling circle (RC), satellites, short interspersed nuclear elements (SINEs) and retroposons. The middle pie chart represent the proportion of each repeat class in the genome: none (94.4%), simple (micro-satellites) (4.11%), low complexity (0.655%), DNA (0.419%), unknown (0.161%), LTRs (0.110%), LINEs (0.052%), RNA (0.027%), RC (0.019%), satellites (0.010%), retroposons (0.005%), SINEs (0.004%).

used for the initial scaffolding assemblies, followed by mapping of the Illumina short reads onto these scaffolds, thus increasing quality of the assembled sequence while preserving whole chromosome integrity. Final polishing, reordering and reorienting of chromosomes, along with masking and classifying of repeat regions, was guided by the most closely related reference genome for each species. Finished genome annotation was both evidence-based and *ab initio*.

Figure 1 summarises data sizes and total yield per sample. The total sequencing data file size for all samples was 139.33 Gigabytes, yielding 58.70 GigaBases of sequence data from 23.71 GigaReads. Figure 2 summarises our analysis workflow. This workflow generated four main outputs for each assembly: genome, proteome, and transcriptome files in FASTA format, and a General Feature Format file (GFF) that contains the coordinates for all proteins and transcripts in the assembly.

Methods

Sample collection, sequencing and software. From the parasite cryobank at Lancaster University, we selected six samples of the species listed above without publicly available reference genomes. Table 1 gives details for strains, isolates, BioSample and BioProject accessions^{17–28}. Illumina HiSeq 4000 and MiSeq sequencing was contracted to BGI Genomics and Aberystwyth University. Nanopore sequencing was performed in-house using MinION FLO-MIN106 flow cells with SQK-LSK109 ligation sequencing protocol. Throughout the text we provide literature citations to software where available. Links to both published and unpublished software used are provided in Table 2. We created public GitHub and Zenodo repositories for the analysis pipeline^{29,30}.

Genome assembly. *De novo* assemblies were performed with Nanopore MinION long reads using Flye³¹. Due to the low quality scores in Nanopore long reads, we mapped high quality Illumina short reads onto the assemblies and created corrected consensus sequences using minimap²³² and SAMtools³³. The consensus sequence was scanned for any contamination or any sequence of vector origin by BLAST+³⁴ on the UniVec database³⁵. Finally, a polishing step was done to minimise gaps using Pilon³⁶.

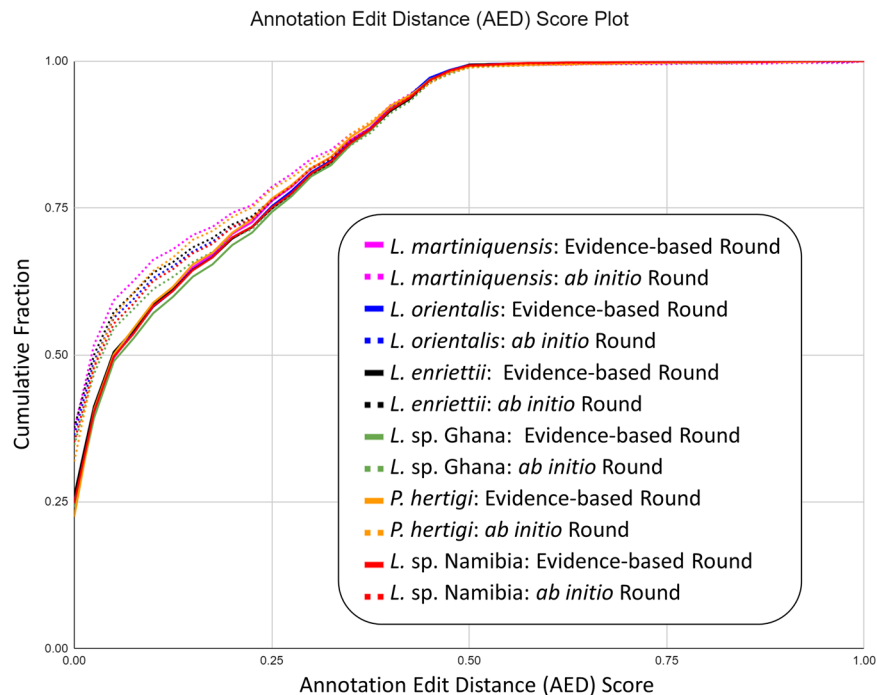


Fig. 5 Annotation Edit Distance (AED) score (x-axis) line plot for all assembly annotation rounds: evidence-based (solid line) and *ab initio* (dotted line). Y-axis represents the genome cumulative percentages.

Chromosome verification. For all chromosomes of each polished genome, we then ran BLAST+ (parameters: -max_target_seqs. 1 -max_hsps 1) against all TriTrypDB³⁷ release-47 genomes. The output for each genome was then visualized using wordcloud to suggest the closest relative among TriTrypDB genomes³⁸. Then, synteny was plotted for each genome by aligning each of its chromosomes with the corresponding chromosomes of its wordcloud-predicted closest relative, using MUMmer³⁹ (Fig. 3). This confirmed that the order and orientation of the chromosomes of each genome was equivalent to those of its closest TriTrypDB genome. Completion was then achieved by sorting and removing any duplicate scaffolds or contigs using funannotate⁴⁰, followed by a final quality check using Genome Assembly Annotation Service (GAAS).

Repetitive element annotation. We identified and classified repeat regions in the polished assemblies using RepeatModeller and TEclass⁴¹. Then, we generated a stratified genome-wide repeat plot for each assembly³⁸ (see also *L. martiniquensis* example in Fig. 4) to assist the decision of which repeats to mask, using RepeatMasker.

Gene prediction and functional annotation. After repeat masking, we annotated the assemblies using the MAKER2⁴² annotation pipeline over two rounds: 1) an evidence-based annotation round using EST, mRNA-seq and protein homology evidence from TriTrypDB release-47 along with our repeat-masking output, 2) an *ab initio* round using AUGUSTUS⁴³, with the pre-trained *L. tarentolae* as the model organism. After each round, Annotation Edit Distance (AED) scores were calculated and plotted (Fig. 5). We calculated brief statistics for each round, e.g. the number of genes and other features, using Genometools⁴⁴ and AGAT⁴⁵. After completion of all annotation rounds, we assigned functional annotations from the Uniprot⁴⁶ and Pfam⁴⁷ databases using BLAST+ and InterProScan⁴⁸.

Analysis pipeline. To make sure that all assemblies and annotations are reproducible by future investigators, the entire process from obtaining the SRAs^{49–91} to the annotation assignments^{92–97} has been made available²⁹ using Snakemake⁹⁸. This Snakemake pipeline ought to be easily adaptable to the sequencing of further similar parasite genomes, throughout the parasitology community³⁰.

Data Records

Table 3 details the sequencing output. Short and long reads were deposited in the NCBI Sequence Read Archive (SRA)^{49–91}. Six BioProjects^{23–28} and six BioSamples^{17–22} were also created at NCBI. The assembled genomes were deposited at NCBI Assembly^{99–104}. Additional files containing raw reads quality reports^{105–110}, mapped reads^{111–116}, classified repeated sequences^{117–122} and functional annotations^{92–97} were deposited at Lancaster University electronic data archive.

species	Sequencing Platforms	SRA Accession	Number of Reads (GigaReads)	Bases (GigaBase)	File size (Gigabyte)
<i>L. (Mundinia) martiniquensis</i>	Illumina HiSeq 4000	SRR13558784	0.783	1.182	2.981
		SRR13558792	1.089	1.644	4.151
	Illumina MiSeq	SRR13558785	0.446	1.327	3.003
	Nanopore MinION	SRR13558786	0.071	3.634	7.323
		SRR13558788	0.006	0.321	0.647
		SRR13558790	0.004	0.468	0.940
<i>L. (Mundinia) orientalis</i>	Illumina HiSeq 2500	SRR13558774	1.579	1.437	4.843
		SRR13558775	0.618	0.563	1.894
		SRR13558776	1.560	1.420	4.786
		SRR13558777	0.636	0.578	1.947
	Illumina HiSeq 4000	SRR13558779	1.079	1.629	4.112
		SRR13558780	1.406	2.123	5.361
	Illumina MiSeq	SRR13558781	0.383	1.135	2.568
	Nanopore MinION	SRR13558782	0.054	3.357	6.756
<i>L. (Mundinia) enriettii</i>	Illumina HiSeq 4000	SRR13558795	0.879	1.328	3.350
		SRR13558796	1.214	1.834	4.630
	Illumina MiSeq	SRR13558797	0.506	1.494	3.385
	Nanopore MinION	SRR13558798	0.072	4.365	8.786
<i>L. (Mundinia) sp. Ghana</i>	Illumina HiSeq 2500	SRR13558800	1.228	1.117	3.765
		SRR13558801	0.684	0.623	2.096
	Illumina HiSeq 4000	SRR13558802	1.006	1.519	3.833
		SRR13558803	1.407	2.124	5.365
	Illumina MiSeq	SRR13558804	0.520	1.549	3.505
	Nanopore MinION	SRR13558805	0.077	5.390	10.840
<i>L. (Mundinia) sp. Namibia</i>	Illumina HiSeq 4000	SRR13558764	0.527	1.567	3.546
		SRR13558765	0.985	1.487	3.753
	Illumina MiSeq	SRR13558766	1.347	2.034	5.136
	Nanopore MinION	SRR13558767	0.068	4.377	8.807
<i>Porcisia hertigi</i>	Illumina HiSeq 4000	SRR13558754	0.929	1.403	3.540
		SRR13558755	1.409	2.128	5.374
	Illumina MiSeq	SRR13558756	0.379	1.123	2.541
	Nanopore MinION	SRR13558757	0.019	1.364	2.742
Grand Total			23.708	58.698	139.327

Table 3. Details of reads, bases and file sizes.

Technical Validation

Genomic DNA integrity. Genomic DNA was extracted using Trizol (Invitrogen) and quantified using Qubit® dsDNA HS Assay Kits (ThermoFisher Scientific) prior to sequencing. Concentrations ranged between 68.2 and 120 ng/μL. For consistency, we used the same extracted DNA for all three sequencing platforms (Nanopore MinION, Illumina HiSeq 4000 and MiSeq). Furthermore, we assessed the gDNA high molecular weight using N50 estimates of MinION long reads which were ranged between 12.07 and 22.92 kilobases.

Contamination screening. We scanned all assemblies for any contamination or any sequence of vector origin by first building a UniVec Database and then using BLAST+. All contaminants were found either at the beginning or at the end of contigs and then deleted. No contaminants affected assembly integrity.

Quality of short and long raw sequence reads. We used FastQC to check the sequence quality of Illumina short reads sequences and pycoQC to check the Nanopore long reads sequence quality. We used MultiQC¹²³ to output all sequence quality scores in one interactive report^{105–110}.

Assembly validation. Since the analysis took many steps to finish, quality checks were introduced between each step. Some checks were focused on completeness, for instance using BUSCO¹²⁴ as a benchmark for the presence of expected universal single-copy orthologues. Other checks focussed on the correct order and orientation of the chromosomes, for instance MUMmer alignment to find synteny between assemblies and other *Leishmania* genomes. Yet further checks focussed on the accuracy and precision of annotation, for instance using Annotation Edit Distance score (AED) in MAKER2 (Fig. 5). We checked reproducibility of the assemblies and annotations using Snakemake.

Code availability

The Snakemake analyses pipeline was deposited at GitHub and Zenodo repositories^{29,30}. Links to software used as well as relevant conda and docker containers are given in Table 2.

Received: 7 May 2021; Accepted: 9 August 2021;

Published online: 06 September 2021

References

1. Steverding, D. The history of leishmaniasis. *Parasit Vectors* **10**, 82–91 (2017).
2. Maroli, M., Feliciangeli, M. D., Bichaud, L., Charrel, R. N. & Gradoni, L. Phlebotomine sandflies and the spreading of leishmaniasis and other diseases of public health concern. *Med Vet Entomol* **27**, 123–147 (2013).
3. Zijlstra, E. E. PKDL and other dermal lesions in HIV co-infected patients with Leishmaniasis: review of clinical presentation in relation to immune responses. *PLoS Negl Trop Dis* **8**, e3258 (2014).
4. Al-Salem, W., Herricks, J. R. & Hotez, P. J. A review of visceral leishmaniasis during the conflict in South Sudan and the consequences for East African countries. *Parasit Vectors* **9**, 460–470 (2016).
5. Burza, S., Croft, S. L. & Boelaert, M. Leishmaniasis. *Lancet* **392**, 951–970 (2018).
6. Desbois, N., Pratlong, F., Quist, D. & Dedet, J. P. *Leishmania (Leishmania) martiniquensis* n. sp. (Kinetoplastida: Trypanosomatidae), description of the parasite responsible for cutaneous leishmaniasis in Martinique Island (French West Indies). *Parasite* **21**, 12–15 (2014).
7. Jariyapan, N. *et al.* *Leishmania (Mundinia) orientalis* n. sp. (Trypanosomatidae), a parasite from Thailand responsible for localised cutaneous leishmaniasis. *Parasit Vectors* **11**, 351–359 (2018).
8. Kwakye-Nuako, G. *et al.* First isolation of a new species of *Leishmania* responsible for human cutaneous leishmaniasis in Ghana and classification in the *Leishmania enriettii* complex. *Int J Parasitol* **45**, 679–684 (2015).
9. Lobsiger, L. *et al.* An autochthonous case of cutaneous bovine leishmaniasis in Switzerland. *Vet Parasitol* **169**, 408–414 (2010).
10. Muller, N. *et al.* Occurrence of *Leishmania* sp. in cutaneous lesions of horses in Central Europe. *Vet Parasitol* **166**, 346–351 (2009).
11. Reuss, S. M. *et al.* Autochthonous *Leishmania siamensis* in horse, Florida, USA. *Emerg Infect Dis* **18**, 1545–1547 (2012).
12. Rose, K. *et al.* Cutaneous leishmaniasis in red kangaroos: isolation and characterisation of the causative organisms. *Int J Parasitol* **34**, 655–664 (2004).
13. Ivens, A. C. & Blackwell, J. M. The *Leishmania* genome comes of age. *Parasitol Today* **15**, 225–231 (1999).
14. Ivens, A. C. *et al.* The genome of the kinetoplastid parasite, *Leishmania major*. *Science* **309**, 436–442 (2005).
15. Albanaz, A. T. S. *et al.* Genome analysis of *Endotrypanum* and *Porcisia* spp., closest phylogenetic relatives of *Leishmania*, highlights the role of amastins in shaping pathogenicity. *Genes (Basel)* **12**, 444–463 (2021).
16. Espinosa, O. A., Serrano, M. G., Camargo, E. P., Teixeira, M. M. G. & Shaw, J. J. An appraisal of the taxonomy and nomenclature of trypanosomatids presently classified as *Leishmania* and *Endotrypanum*. *Parasitology* **145**, 430–442 (2018).
17. NCBI BioSample <https://identifiers.org/ncbi/biosample:SAMN17294109> (2021).
18. NCBI BioSample <https://identifiers.org/ncbi/biosample:SAMN17294111> (2021).
19. NCBI BioSample <https://identifiers.org/ncbi/biosample:SAMN17294112> (2021).
20. NCBI BioSample <https://identifiers.org/ncbi/biosample:SAMN17294115> (2021).
21. NCBI BioSample <https://identifiers.org/ncbi/biosample:SAMN17294129> (2021).
22. NCBI BioSample <https://identifiers.org/ncbi/biosample:SAMN17294121> (2021).
23. NCBI BioProject <https://identifiers.org/ncbi/bioproject:PRJNA691531> (2021).
24. NCBI BioProject <https://identifiers.org/ncbi/bioproject:PRJNA691532> (2021).
25. NCBI BioProject <https://identifiers.org/ncbi/bioproject:PRJNA691534> (2021).
26. NCBI BioProject <https://identifiers.org/ncbi/bioproject:PRJNA691536> (2021).
27. NCBI BioProject <https://identifiers.org/ncbi/bioproject:PRJNA689706> (2021).
28. NCBI BioProject <https://identifiers.org/ncbi/bioproject:PRJNA691541> (2021).
29. Almutairi, H. *hatalmutairi/LGAAP*. <https://doi.org/10.5281/zenodo.4663265> (2021).
30. Almutairi, H. *et al.* LGAAP: *Leishmaniinae* Genome Assembly and Annotation Pipeline. *Microbiol Resour Announc* **10**, e0043921 (2021).
31. Kolmogorov, M., Yuan, J., Lin, Y. & Pevzner, P. A. Assembly of long, error-prone reads using repeat graphs. *Nat Biotechnol* **37**, 540–546 (2019).
32. Li, H. Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. *Bioinformatics* **32**, 2103–2110 (2016).
33. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *Gigascience* **10**, giab008. <https://doi.org/10.1093/gigascience/giab008> (2021).
34. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421–428 (2009).
35. NCBI. The UniVec Database. <https://www.ncbi.nlm.nih.gov/tools/vecscreen/univec/> (2016).
36. Walker, B. J. *et al.* Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* **9**, e112963 (2014).
37. Aslett, M. *et al.* TriTrypDB: a functional genomic resource for the Trypanosomatidae. *Nucleic Acids Res* **38**, D457–462 (2010).
38. Almutairi, H. Supplementary materials for chromosome-scale genome sequencing, assembly and annotation of six genomes from subfamily *Leishmaniinae*. *Lancaster University* <https://doi.org/10.17635/lancaster/researchdata/474> (2021).
39. Delcher, A. L., Salzberg, S. L. & Phillippy, A. M. Using MUMmer to identify similar regions in large sequence sets. *Curr Protoc Bioinformatics* Chapter 10: Unit 10.3. <https://doi.org/10.1002/0471250953.bi1003s00> (2003).
40. Palmer, J. & Stajich, J. nextgenusfs/funannotate: funannotate v1.5.3 (Version 1.5.3). *Zenodo*. <https://doi.org/10.5281/zenodo.2604804> (2019).
41. Abrusan, G., Grundmann, N., DeMester, L. & Makalowski, W. TEclass—a tool for automated classification of unknown eukaryotic transposable elements. *Bioinformatics* **25**, 1329–1330 (2009).
42. Holt, C. & Yandell, M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics* **12**, 491 (2011).
43. Hoff, K. J. & Stanke, M. Predicting genes in single genomes with AUGUSTUS. *Curr Protoc Bioinformatics* **65**, e57 (2019).
44. Gremme, G., Steinbiss, S. & Kurtz, S. GenomeTools: a comprehensive software library for efficient processing of structured genome annotations. *IEEE/ACM Trans Comput Biol Bioinform* **10**, 645–656 (2013).
45. Dainat, J., Hereñú, D., & Pucholt, P. NBISweden/AGAT: AGAT-v0.7.0 (v0.7.0). *Zenodo*. <https://doi.org/10.5281/zenodo.5036996> (2021).
46. UniProt, C. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res* **49**, D480–D489 (2021).
47. Mistry, J. *et al.* Pfam: The protein families database in 2021. *Nucleic Acids Res* **49**, D412–D419 (2021).
48. Jones, P. *et al.* InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).
49. NCBI Sequence Read Archive <https://identifiers.org/ncbi/insdc.sra:SRX9957074> (2021).
50. NCBI Sequence Read Archive <https://identifiers.org/ncbi/insdc.sra:SRX9957073> (2021).
51. NCBI Sequence Read Archive <https://identifiers.org/ncbi/insdc.sra:SRX9957072> (2021).
52. NCBI Sequence Read Archive <https://identifiers.org/ncbi/insdc.sra:SRX9957071> (2021).

53. NCBI Sequence Read Archive <https://identifiers.org/ncbi/insdc.sra:SRX9957070> (2021).
54. NCBI Sequence Read Archive <https://identifiers.org/ncbi/insdc.sra:SRX9957069> (2021).
55. NCBI Sequence Read Archive <https://identifiers.org/ncbi/insdc.sra:SRX9957068> (2021).
56. NCBI Sequence Read Archive <https://identifiers.org/ncbi/insdc.sra:SRX9957067> (2021).
57. NCBI Sequence Read Archive <https://identifiers.org/ncbi/insdc.sra:SRX9957066> (2021).
58. NCBI Sequence Read Archive <https://identifiers.org/ncbi/insdc.sra:SRX9957065> (2021).
59. NCBI Sequence Read Archive <https://identifiers.org/ncbi/insdc.sra:SRX9957064> (2021).
60. NCBI Sequence Read Archive <https://identifiers.org/ncbi/insdc.sra:SRX9957063> (2021).
61. NCBI Sequence Read Archive <https://identifiers.org/ncbi/insdc.sra:SRX9957062> (2021).
62. NCBI Sequence Read Archive <https://identifiers.org/ncbi/insdc.sra:SRX9957061> (2021).
63. NCBI Sequence Read Archive <https://identifiers.org/ncbi/insdc.sra:SRX9957060> (2021).
64. NCBI Sequence Read Archive <https://identifiers.org/ncbi/insdc.sra:SRX9957059> (2021).
65. NCBI Sequence Read Archive <https://identifiers.org/ncbi/insdc.sra:SRX9957058> (2021).
66. NCBI Sequence Read Archive <https://identifiers.org/ncbi/insdc.sra:SRX9957057> (2021).
67. NCBI Sequence Read Archive <https://identifiers.org/ncbi/insdc.sra:SRX9957056> (2021).
68. NCBI Sequence Read Archive <https://identifiers.org/ncbi/insdc.sra:SRX9957055> (2021).
69. NCBI Sequence Read Archive <https://identifiers.org/ncbi/insdc.sra:SRX9957054> (2021).
70. NCBI Sequence Read Archive <https://identifiers.org/ncbi/insdc.sra:SRX9957079> (2021).
71. NCBI Sequence Read Archive <https://identifiers.org/ncbi/insdc.sra:SRX9957078> (2021).
72. NCBI Sequence Read Archive <https://identifiers.org/ncbi/insdc.sra:SRX9957077> (2021).
73. NCBI Sequence Read Archive <https://identifiers.org/ncbi/insdc.sra:SRX9957076> (2021).
74. NCBI Sequence Read Archive <https://identifiers.org/ncbi/insdc.sra:SRX9957075> (2021).
75. NCBI Sequence Read Archive <https://identifiers.org/ncbi/insdc.sra:SRX9957086> (2021).
76. NCBI Sequence Read Archive <https://identifiers.org/ncbi/insdc.sra:SRX9957085> (2021).
77. NCBI Sequence Read Archive <https://identifiers.org/ncbi/insdc.sra:SRX9957084> (2021).
78. NCBI Sequence Read Archive <https://identifiers.org/ncbi/insdc.sra:SRX9957083> (2021).
79. NCBI Sequence Read Archive <https://identifiers.org/ncbi/insdc.sra:SRX9957082> (2021).
80. NCBI Sequence Read Archive <https://identifiers.org/ncbi/insdc.sra:SRX9957081> (2021).
81. NCBI Sequence Read Archive <https://identifiers.org/ncbi/insdc.sra:SRX9957080> (2021).
82. NCBI Sequence Read Archive <https://identifiers.org/ncbi/insdc.sra:SRX9957038> (2021).
83. NCBI Sequence Read Archive <https://identifiers.org/ncbi/insdc.sra:SRX9957037> (2021).
84. NCBI Sequence Read Archive <https://identifiers.org/ncbi/insdc.sra:SRX9957036> (2021).
85. NCBI Sequence Read Archive <https://identifiers.org/ncbi/insdc.sra:SRX9957035> (2021).
86. NCBI Sequence Read Archive <https://identifiers.org/ncbi/insdc.sra:SRX9957034> (2021).
87. NCBI Sequence Read Archive <https://identifiers.org/ncbi/insdc.sra:SRX9957048> (2021).
88. NCBI Sequence Read Archive <https://identifiers.org/ncbi/insdc.sra:SRX9957047> (2021).
89. NCBI Sequence Read Archive <https://identifiers.org/ncbi/insdc.sra:SRX9957046> (2021).
90. NCBI Sequence Read Archive <https://identifiers.org/ncbi/insdc.sra:SRX9957045> (2021).
91. NCBI Sequence Read Archive <https://identifiers.org/ncbi/insdc.sra:SRX9957044> (2021).
92. Almutairi, H. L. (*Mundinia*) *martiniquensis*: functional annotations. *Lancaster University* <https://doi.org/10.17635/lancaster/researchdata/446> (2021).
93. Almutairi, H. L. (*Mundinia*) *orientalis*: functional annotations. *Lancaster University* <https://doi.org/10.17635/lancaster/researchdata/449> (2021).
94. Almutairi, H. L. (*Mundinia*) *enriettii*: functional annotations. *Lancaster University* <https://doi.org/10.17635/lancaster/researchdata/452> (2021).
95. Almutairi, H. L. (*Mundinia*) sp. Ghana: functional annotations. *Lancaster University* <https://doi.org/10.17635/lancaster/researchdata/455> (2021).
96. Almutairi, H. L. (*Mundinia*) sp. Namibia: functional annotations. *Lancaster University* <https://doi.org/10.17635/lancaster/researchdata/458> (2021).
97. Almutairi, H. *Porcisia hertigi*: functional annotations. *Lancaster University* <https://doi.org/10.17635/lancaster/researchdata/461> (2021).
98. Mölder, F. *et al.* Sustainable data analysis with Snakemake. *F1000Research* **10**, <https://f1000research.com/articles/10-33/v2> (2021).
99. NCBI Assembly https://identifiers.org/insdc.gca:GCA_017916325.1 (2021).
100. NCBI Assembly https://identifiers.org/insdc.gca:GCA_017916335.1 (2021).
101. NCBI Assembly https://identifiers.org/insdc.gca:GCA_017916305.1 (2021).
102. NCBI Assembly https://identifiers.org/insdc.gca:GCA_017918215.1 (2021).
103. NCBI Assembly https://identifiers.org/insdc.gca:GCA_017918225.1 (2021).
104. NCBI Assembly https://identifiers.org/insdc.gca:GCA_017918235.1 (2021).
105. Almutairi, H. L. (*Mundinia*) *martiniquensis* raw reads quality reports. *Lancaster University* <https://doi.org/10.17635/lancaster/researchdata/437> (2021).
106. Almutairi, H. *Leishmania* (*Mundinia*) *orientalis* raw reads quality reports. *Lancaster University* <https://doi.org/10.17635/lancaster/researchdata/438> (2021).
107. Almutairi, H. *Leishmania* (*Mundinia*) *enriettii* raw reads quality reports. *Lancaster University* <https://doi.org/10.17635/lancaster/researchdata/439> (2021).
108. Almutairi, H. *Leishmania* (*Mundinia*) sp. Ghana raw reads quality reports. *Lancaster University* <https://doi.org/10.17635/lancaster/researchdata/440> (2021).
109. Almutairi, H. *Leishmania* (*Mundinia*) sp. Namibia raw reads quality reports. *Lancaster University* <https://doi.org/10.17635/lancaster/researchdata/441> (2021).
110. Almutairi, H. *Porcisia hertigi* raw reads quality reports. *Lancaster University* <https://doi.org/10.17635/lancaster/researchdata/442> (2021).
111. Almutairi, H. L. (*Mundinia*) *martiniquensis*: mapped reads in SAM and BAM format. *Lancaster University* <https://doi.org/10.17635/lancaster/researchdata/444> (2021).
112. Almutairi, H. L. (*Mundinia*) *orientalis*: mapped reads in SAM and BAM format. *Lancaster University* <https://doi.org/10.17635/lancaster/researchdata/447> (2021).
113. Almutairi, H. L. (*Mundinia*) *enriettii*: mapped reads in SAM and BAM format. *Lancaster University* <https://doi.org/10.17635/lancaster/researchdata/450> (2021).
114. Almutairi, H. L. (*Mundinia*) sp. Ghana: mapped reads in SAM and BAM format. *Lancaster University* <https://doi.org/10.17635/lancaster/researchdata/453> (2021).
115. Almutairi, H. L. (*Mundinia*) sp. Namibia: mapped reads in SAM and BAM format. *Lancaster University* <https://doi.org/10.17635/lancaster/researchdata/456> (2021).
116. Almutairi, H. *Porcisia hertigi*: mapped reads in SAM and BAM format. *Lancaster University* <https://doi.org/10.17635/lancaster/researchdata/459> (2021).

117. Almutairi, H. L. (*Mundinia*) *martiniquensis*: classified repeated sequences. *Lancaster University* <https://doi.org/10.17635/lancaster/researchdata/445> (2021).
118. Almutairi, H. L. (*Mundinia*) *orientalis*: classified repeated sequences. *Lancaster University* <https://doi.org/10.17635/lancaster/researchdata/448> (2021).
119. Almutairi, H. L. (*Mundinia*) *enriettii*: classified repeated sequences. *Lancaster University* <https://doi.org/10.17635/lancaster/researchdata/451> (2021).
120. Almutairi, H. L. (*Mundinia*) sp. Ghana: classified repeated sequences. *Lancaster University* <https://doi.org/10.17635/lancaster/researchdata/454> (2021).
121. Almutairi, H. L. (*Mundinia*) sp. Namibia: classified repeated sequences. *Lancaster University* <https://doi.org/10.17635/lancaster/researchdata/457> (2021).
122. Almutairi, H. *Porcisia hertigi*: classified repeated sequences. *Lancaster University* <https://doi.org/10.17635/lancaster/researchdata/460> (2021).
123. Ewels, P., Magnusson, M., Lundin, S. & Kaller, M. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* **32**, 3047–3048 (2016).
124. Seppey, M., Manni, M. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness. *Methods Mol Biol* **1962**, 227–245 (2019).

Acknowledgements

We thank the Ministry of Health and Public Health Authority of Saudi Arabia for funding.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to D.G.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

The Creative Commons Public Domain Dedication waiver <http://creativecommons.org/publicdomain/zero/1.0/> applies to the metadata files associated with this article.

© The Author(s) 2021