Data in Brief 46 (2023) 108811

Contents lists available at ScienceDirect

Data in Brief

journal homepage: www.elsevier.com/locate/dib

Data Article

Dataset of dual RNA-seq mapping in visceral leishmaniasis: Inquiry on parasite transcripts in human blood transcriptome upon Leishmania infantum infection

Ellen Gomes[a], Luana Aparecida Rogerio[a], Nayore Tamie Takamiya[a], Caroline Torres[a], João Santana da Silva[b], Roque Pacheco Almeida[c], Sandra Regina Maruyama[a,*]

[a] Department of Genetics and Evolution, Center for Biological Sciences and Health, Federal University of São Carlos, São Carlos, Brazil
[b] Fiocruz-Bi-Institutional Translational Medicine Platform, Ribeirão Preto, Brazil
[c] Department of Medicine, University Hospital-Empresa Brasileira de Serviços Hospitalares (EBSERH), Federal University of Sergipe, Aracaju, Brazil

ARTICLE INFO

Article history:
Received 27 September 2022
Revised 16 November 2022
Accepted 2 December 2022
Available online 8 December 2022

Dataset link: Dual mapping of RNA-seq samples from blood of visceral leishmaniasis patients. (Original data)

Keywords:
Dual RNA-seq
Host/parasite interaction
Leishmania
Leishmaniasis
blood transcriptomics

ABSTRACT

This dataset is related to the article "Insight Into the Long Noncoding RNA and mRNA Coexpression Profile in the Human Blood Transcriptome Upon Leishmania infantum Infection" by S.R. Maruyama, C.A. Fuzo, A.E.R. Oliveira, L.A. Rogerio, N.T. Takamiya, G. Pessenda, E.V. de Melo, A.M. da Silva, A.R. Jesus, V. Carregaro, H.I. Nakaya, R.P. Almeida and J.S. da Silva. Frontiers in Immunology, 2022. Through the reuse of raw sequencing data, we generated original dataset by performing a dual RNA-seq mapping procedure to survey the parasite transcripts found in RNA-seq samples from blood of visceral leishmaniasis patients. Diseased patients with active infection displayed the highest number of reads mapped to *L. infantum* genome. Even after six months later of the treatment, when the patients were considered cured, parasite reads were still detected. Parasite reads were also detected in asymptomatic individuals. The original dual RNA-seq align-

* Corresponding author.
E-mail address: sandrarcm@ufscar.br (S.R. Maruyama).
Social media: @Nayore_Takamiya (N.T. Takamiya), @sandra_maruyama (S.R. Maruyama)

https://doi.org/10.1016/j.dib.2022.108811

ment read count data provided here can be further explored to evaluate either host or parasite transcripts.

Specifications Table

Subject	Biological science
Specific subject area	*Bioinformatics and Transcriptomics*
Type of data	Tables, spreadsheets, text files
How the data were acquired	The mRNA-seq data from blood of nineteen individuals were retrieved from the gene expression study under accession number E-MTAB-11047 available on ArrayExpress repository.
Data format	Raw sequence files (re-used from E-MTAB-11047) Raw read count files (original) Analyzed (original) Filtered (original)
Description of data collection	Raw fastq files of samples classified according to the *L. infantum* infection state as PD0 (active disease state, before treatment), PD180 (cured disease state, 180 days after treatment) or A (asymptomatic individuals, no clinical signs) were analyzed. Reads were checked for quality control and trimmed. High quality reads were mapped to the concatenated human and *Leishmania infantum* genomes using STAR Aligner. The option "quantMode GeneCounts" within STAR was used to perform reads counting. STAR Output tables with read counts were analyzed focused on alignment in genome parasite to produce a survey for parasite gene transcription dataset (original data).
Data source location	The raw mRNA-seq data are available on ArrayExpress, E-MTAB-11047, 2022, https://www.ebi.ac.uk/biostudies/arrayexpress/studies/E-MTAB-11047. SANDRA MARUYAMA (submitter). "Bulk mRNA-seq of blood samples of visceral leishmaniasis patients during active disease and after treatment (cured), asymptomatic individuals and healthy uninfected control volunteers." Accessed 26 May 2022 (Re-used). The original data (raw, processed, filtered and analyzed) are available at Mendeley Data Repository: "Dual mapping of RNA-seq samples from blood of visceral leishmaniasis patients". DOI:10.17632/2xwjzjzvyn.1
Data accessibility	Raw sequencing data (fastq files) at Functional Genomics Repository (ArrayExpress) under accession number E-MTAB-11047: https://www.ebi.ac.uk/biostudies/arrayexpress/studies/E-MTAB-11047 Raw read count files and processed/analyzed data of dual mapping procedure: Mendeley Data Repository, Version 1 available at https://data.mendeley.com/datasets/2xwjzjzvyn
Related research article	S.R. Maruyama, C.A. Fuzo, A.E.R. Oliveira, L.A. Rogerio, N.T. Takamiya, G. Pessenda, E.V. de Melo, A.M. da Silva, A.R. Jesus, V. Carregaro, H.I. Nakaya, R.P. Almeida and J.S. da Silva (2022). Insight Into the Long Noncoding RNA and mRNA Coexpression Profile in the Human Blood Transcriptome Upon *Leishmania infantum* Infection. Front. Immunol. 13:784463. doi:10.3389/fimmu.2022.784463 [1].

Value of the Data

- The dataset of parasite transcripts in different outcomes of *Leishmania infantum* infection can guiding further studies for biomarkers prediction and validation.
- Mining RNA-seq data of infected hosts to search reads from pathogens can reveal information about gene transcription of *Leishmania infantum* parasites.
- Research field of parasite/host interaction in Leishmaniases can explore the data to gather insights from a dynamic perspective of the infection and the disease.

1. Objective

Leishmania infections in humans target host tissues such as skin, spleen, liver and bone marrow. These infections are not characterized by parasitemia, except for patients co-infected with HIV, who present a higher load of parasites in blood [2]. *Leishmania* parasites are present in peripheral blood, however due to the low load, they can be found in blood samples only through sensitive nucleic acid detection methods such as qPCR (quantitative **p**olymerase **c**hain **r**eaction) [3]. Elsewhere, we analyzed blood transcriptomes of visceral leishmaniasis patients infected with *L. infantum* compared to the non-diseased groups: asymptomatics (no clinical signs of VL, but present positive antibody detection for *L. infantum*) cured (six months after successful treatment) and uninfected health controls [1]. Because the data were obtained through RNA sequencing (RNA-seq), a burning question arises: Are *Leishmania* transcripts detected in the RNA-seq data of blood from VL patients?

In order to survey the parasite transcripts found in blood from patients with active infection, six months after treatment (cured), as well as in asymptomatic individuals, here our aim was to perform a dual RNA-seq mapping analysis through re-using of previous RNA-seq data [1].

2. Data Description

The dataset presented here is a combination of data mining and original data production. It was built through reusing raw sequencing data (fastq files) from Maruyama et al. study [1] (available at https://www.ebi.ac.uk/biostudies/arrayexpress/studies/E-MTAB-11047) followed by an original analysis through dual RNA-seq mapping (available at https://data.mendeley.com/datasets/2xwjzjzvyn). Approximately 79% of reads mapped uniquely to the human/*L. infantum* concatenated genome. The survey for parasite transcription data yielded an original dataset comprised by a total of 2,601 reads mapped to 1,419 *L. infantum* genes distributed across the 36 chromosomes of parasite genome (Supplementary Table 1). The data are summarized in Table 1 and Table 2 and described as follow.

Table 1: all samples from PD0 group (patients with active disease before treatment) presented reads aligned to *L. infantum* genome totaling 1,597 reads aligned to 1,173 genes ($\sim$ 1.4 reads/transcript). In the PD180 group (same patients after six months of treatment), where individuals were considered clinically cured, reads were detected in much lower number than PD0 group, but still 6 out of 10 samples presented read alignment (P11_D180, P20_D180, P31_D180, P42_D180, P43_D180 and P46_D180). Among diseased state/active infection samples, P6D0 presented the highest read count (638) and six months later (P6D180) no parasite read was found. Except for P31 individual, all other patients presented important reduction of parasite read counts after six months of the treatment (PD0 samples compared to PD180 samples). Parasite reads were also detected in all samples from asymptomatic group, comprising 265 reads mapped to 263 genes ($\sim$ 1 read/transcript).

Table 2: none parasite gene presented read mapping across all samples (i.e., shared among all gene lists). Thirteen *L. infantum* genes presented read alignment at least in three samples within a group. Most of them (8) are annotated as “hypothetical protein – conserved”, i.e., with no function described. The others presented role in different aspects of cellular metabolism, such as protein synthesis (LINF_320009600 and LINF_200021000), calcium signaling (LINF_160007900), oxidation (LINF_340027900) and microtubule motor activity (LINF_130022000). Seven out of 13 genes were detected in RNA-seq of PD0 group, i.e., in human blood upon *L. infantum* infection (LINF_260017300, LINF_320009600, LINF_030011500, LINF_160007900, LINF_060013700, LINF_200021000 and LINF_340027900).

Table 1
Summary of read counts and mapped reads to host and parasite concatenated genomes.

Group	Samples	N° of reads[a]	% of mapped reads[b]	N° of reads mapped to *L. infantum*	N° of *L. infantum* mapped genes	N° of reads mapped to human	N° of human mapped genes
PD0	P6D0	34,881,495	73.09%	638	536	25,497,422	25,157
	P8D0	36,915,299	71.80%	3	3	26,508,362	23,301
	P11D0	24,646,435	80.02%	162	143	19,722,597	25,170
	P20D0	27,840,339	79.67%	357	223	22,182,786	24,097
	P24D0	32,646,169	77.03%	28	23	25,149,249	26,419
	P29D0	30,364,552	82.36%	5	5	25,011,003	24,222
	P31D0	26,775,996	79.14%	19	19	21,191,752	24,367
	P42D0	27,861,819	78.87%	230	121	21,976,994	20,950
	P43D0	31,078,212	79.39%	47	42	24,674,666	24,047
	P46D0	26,935,489	76.33%	108	58	21,004,794	23,832
PD180	P6D180	35,710,907	76.25%	0	0	27,230,094	24,160
	P8D180	29,459,895	83.16%	0	0	29,499,521	23,401
	P11D180	27,208,817	80.31%	28	28	21,853,829	25,431
	P20D180	31,972,955	77.96%	45	45	51,663,789	49,213
	P24D180	28,259,473	78.56%	0	0	22,201,489	25,650
	P29D180	32,539,028	77.94%	0	0	25,364,093	26,175
	P3D180	33,901,027	72.51%	25	22	24,584,175	23,957
	P42D180	24,511,867	79.78%	36	36	19,556,829	23,298
	P43D180	26,096,811	80.48%	28	28	21,004,794	23,832
	P46D180	26,478,354	79.57%	37	36	21,070,772	23,905
Asymptomatic	A1	26,333,879	82.62%	15	15	21,757,658	23,654
	A6	28,348,081	83.51%	25	25	23,674,070	23,657
	A7	22,745,866	77.11%	56	54	17,540,399	25,243
	A8	27,372,410	81.45%	13	13	22,295,025	24,782
	A9	33,194,383	81.04%	18	18	26,902,178	24,815
	A11	23,480,413	73.73%	43	43	17,314,267	25,342
	A13	33,675,198	83.75%	26	26	28,206,021	24,477
	A17	27,350,834	79.61%	29	29	21,775,748	25,387
	A24	26,352,907	80.13%	40	40	21,116,934	24,304

[a] After trimming and quality control procedure.
[b] Uniquely mapped reads.

Table 2
Parasite genes identified in RNA-seq data of blood from visceral patients (PD0), cured VL (PD180) and asymptomatic individuals (A). Only genes found in three or more samples are displayed (icon ∩: intersection).

Group	Samples	Gene ID	Average N° of reads	Chromosome Location	Gene Description	Gene Type
PD0s	P20D0 ∩ P42D0 ∩ P43D0 ∩ P46D0	LINF_260017300	2.75	26	hypothetical protein - conserved	protein coding gene
	P11D0 ∩ P20D0 ∩ P6D0	LINF_320009600	1.00	32	40S ribosomal protein S2	protein coding gene
	P20D0 ∩ P43D0 ∩	LINF_030011500	1.00	3	hypothetical protein - conserved	protein coding gene
	P6D0	LINF_160007900	1.00	16	inositol 1 -4 -5-trisphosphate receptor - putative	protein coding gene
	P31D0 ∩ P46D0 ∩ P6D0	LINF_060013700	4.50	6	hypothetical protein - conserved	protein coding gene
	P11D0 ∩ P20D0 ∩ P43D0	LINF_200021000	2.00	20	40S ribosomal protein S11 - putative	protein coding gene

(*continued on next page*)

Table 2 (*continued*)

Group	Samples	Gene ID	Average N° of reads	Chromosome Location	Gene Description	Gene Type
PD180s	P11D180 ∩ P20D180 ∩ P42D180	LINF_340027900	1.00	32	Kelch motif/Galactose oxidase - central domain containing protein - putative	protein coding gene
	P11D180 ∩ P20D180 ∩ P43D180	LINF_320011300	1.00	32	hypothetical protein - conserved	protein coding gene
	P20D180 ∩ P43D180 ∩	LINF_280021400	1.00	28	hypothetical protein - conserved	protein coding gene
	P46D180	LINF_260017300	1.33	26	hypothetical protein - conserved	protein coding gene
Asymptomatics	A1 ∩ A11 ∩ A6 A7	LINF_060013700	1.25	6	hypothetical protein - conserved	protein coding gene
	A17 ∩ A6 ∩ A7	LINF_320011300	1.00	32	hypothetical protein - conserved	protein coding gene
	A13 ∩ A17 ∩ A8	LINF_130022000	1.00	13	dynein heavy chain - putative	protein coding gene

3. Experimental Design Materials and Methods

Here, we re-used raw fastq files from E-MTAB-11047 study [1] and presented original data from the dual RNA-seq mapping approach performed in the blood RNA-seq data of visceral leishmaniasis patients before (active disease state, PD0 group, n=10, being 5 women and 5 men) and after six months of the treatment (cured disease state, PD180 group, n=10, the same 5 women and 5 men). Also, the same procedure was performed in samples from asymptomatic individuals (A group, n=9, being 3 women and 6 men). The age in PD0 and PD180 groups ranged from 1 to 51 years-old for women patients and from 10 to 44 years-old for men patients (age mean values 14,5 and 24 years-old, respectively). For asymptomatic group the age ranged from 3 to 30 years-old for women and from 6 to 42 years old (age mean values 12 and 21 years-old respectively).

These 19 individuals were grouped based on the phase of visceral leishmaniasis disease. Patients presenting fever, weight loss, hepatosplenomegaly, and low leukocyte and platelet counts with confirmed visceral leishmaniasis diagnosis by direct observation of *Leishmania* parasites in bone marrow aspirate or positive culture in Novy–MacNeal–Nicolle (NNN) medium and positive rK39 serological test (Kalazar Detect Rapid Test, InBios International Inc. Seattle, WA) was considered as patients with active visceral leishmaniasis disease (PD0). These patients were treated with meglumine antimoniate (Glucantime®) and/or liposomal amphotericin B (AmBisome®) and after 180 days the patients were considered clinically cured (PD180). Healthy individuals who presented normal hematologic indices and neither clinical signs nor symptoms of VL but positive reactions to leishmanial antigens (Montenegro Skin test and rK39 serological test) were considered asymptomatic (A). All VL patients were negative for hepatitis B and C viruses and HIV, and also for bacterial infections or other parasites. Of note, the individuals enrolled in this study are from Sergipe state, located in the Northeast region of Brazil, that is not endemic for Malaria.

A concatenated host/parasite (human/*L. infantum*) reference genome was used for read alignment (dual RNA-seq mapping). The read count output files from STAR aligner (available at Mendeley Data, Reserved DOI:10.17632/2xwjzjzvyn.1) were compiled into one table (Supplementary Table 1) and used to survey specifically the parasite transcripts. To identify the parasite transcripts that were consistently found across the samples, Venn diagrams were performed using samples' gene lists.

3.1. RNA isolation and mRNA sequencing

Peripheral blood samples were collected using BD Vacutainer® tubes for hematologic tests and PAXgene Blood RNA tubes for RNA isolation. Total RNA was extracted from whole blood with the PAXgene Blood RNA Kit followed by globin mRNA depletion using the GLOBINclear™ Human Kit to enhance the samples for RNA from leukocytes. Concentration, purity and integrity were determined using Qubit™ 3.0 Fluorometer with a Qubit™ RNA HS Assay Kit, NanoDrop™ 1000 Spectrophotometer by absorbance measurements (nm) of the 260/280 and 260/230 ratios and Agilent 2100 Bioanalyzer using a Bioanalyzer RNA 6000 Nano assay, respectively. Samples with a RIN $>$ 7 were submitted for RNA sequencing. mRNA sequencing was performed at the Genomics Center of the Laboratory of Animal Biotechnology, ESALQ, University of São Paulo, Piracicaba, Brazil, using Illumina sequencing technology. Polyadenylated cDNA libraries were prepared with 300 ng of RNA depleted from globin mRNAs using the TruSeq® Stranded RNA Sample Preparation Kit. Paired-end sequencing was performed using a HiSeq SBS V4 kit (2 $\times$ 100 or 2 $\times$ 125 reads) in a HiSeq 2500 sequencer, yielding approximately 71 million reads for each mRNA-seq library.

3.2. Data analyses

A total of 29 raw fastq files were used in this study. Illumina sequencing adaptors from raw mRNA-seq data of asymptomatics individuals, visceral leishmaniasis patients before and after treatment were trimmed and the low-quality reads were filtered using Trimmomatic v. 0.39 [4] with the following parameters: LEADING:3 TRAILING:3 SLIDINGWINDOW:4:20 CROP:100 (or CROP:125). An average of 92.85 % of raw paired-end reads were kept during the trimming process. An average of 29 million high quality reads per sample were used for dual mapping procedure.

Genome and annotation indexes were generated through the concatenation of *Leishmania infantum* GCA_900500625.1) and Homo sapiens GRCh38.p13 (available at https://protists.ensembl.org/Leishmania_infantum_gca_900500625/Info/Index) (available at http://www.ensembl.org/Homo_sapiens/Info/Index) files obtained from ENSEMBL database. High quality reads were then aligned using STAR aligner v.2.7.10a [5]. Alignments were run using these combined parasite and human references using ––outFilterMatchNmin 40 to locally align to account for splice leader sequences within parasite transcripts, which may otherwise be discarded [6]. Other STAR alignment parameters as follows: –outFilterMultimapNmax 20 –outFilterMismatchNmax 15 –alignSJoverhangMin 8 –alignMatesGapMax 1000000 ––outFilterMatchNmin 40 –quantMode GeneCounts. The parasite and human counting number of reads per gene were accessed with –quantMode GeneCounts option during the alignment.

The percentage of reads aligned in each genome was calculated using read count files from STAR output (available at Mendeley Data, Reserved DOI:10.17632/2xwjzjzvyn.1). The annotation for *L. infantum* genes was performed using the TriTryp database [7]. The read count data were also used to ascertain the read overlaps across samples per group using Venn diagrams (available at https://bioinformatics.psb.ugent.be/webtools/Venn/). In the INPUT section, the list of gene IDs with read count $>$ 0 per sample was used. This was executed for all samples per group. The Venn diagram output was a list of genes in a text file showing which genes are shared per sample.

Ethics Statements

All procedures performed were approved by the Ethics Committee of the Federal University of Sergipe (CAAE: 04587312.2.0000.0058) according to the recommendations of the Brazilian Human Research Ethics Evaluation System (CEP/CONEP). All subjects or their legal guardians signed an informed consent form previous to the study.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data Availability

Dual mapping of RNA-seq samples from blood of visceral leishmaniasis patients (Original data) (Mendeley Data).

CRediT Author Statement

Ellen Gomes: Methodology, Formal analysis, Visualization, Writing – original draft; **Luana Aparecida Rogerio:** Formal analysis, Visualization, Writing – original draft; **Nayore Tamie Takamiya:** Formal analysis, Visualization; **Caroline Torres:** Formal analysis, Visualization; **João Santana da Silva:** Methodology, Resources, Writing – review & editing; **Roque Pacheco Almeida:** Methodology, Resources, Writing – review & editing; **Sandra Regina Maruyama:** Conceptualization, Methodology, Validation, Formal analysis, Investigation, Resources, Data curation, Writing – original draft, Writing – review & editing, Visualization, Supervision, Project administration, Funding acquisition.

Acknowledgments

We thank Iran Malavazi, Anderson Ferreira Cunha and Felipe Roberti Teixeira for their generous and continuing support. This work was supported by grants from Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP agreements 2016/20258-0 Young Investigator Award to Sandra Regina Maruyama) and 2013/08216-2 (Center for Research in Inflammatory Diseases); from Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES, Finance code 001) and from Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq, grant no. 552721/2011-5). SRM received a fellowship from FAPESP (2017/16328-6). Scholarships from FAPESP were awarded to LAR (2020/14011-8), NTT (2021/12464-8), EG (2022/01525-9) and CT (2021/10358-6).

References

[1] S.R. Maruyama, C.A. Fuzo, A.E.R. Oliveira, L.A. Rogerio, N.T. Takamiya, G. Pessenda, E.V. de Melo, A.M. da Silva, A.R. Jesus, V. Carregaro, H.I. Nakaya, R.P. Almeida, J.S. da Silva, Insight Into the Long Noncoding RNA and mRNA Coexpression Profile in the Human Blood Transcriptome Upon Leishmania infantum Infection, Front. Immunol. 13 (2022) 784463, doi:10.3389/fimmu.2022.784463.

[2] S. Burza, S.L. Croft, M. Boelaert, Leishmaniasis, The Lancet 392 (2018) 951–970, doi:10.1016/S0140-6736(18)31204-2.

[3] S. Moulik, S. Sengupta, M. Chatterjee, Molecular Tracking of the Leishmania Parasite, Front. Cell. Infect. Microbiol. 11 (2021) https://www.frontiersin.org/articles/10.3389/fcimb.2021.623437. accessed August 3, 2022.

[4] USADELLAB.org - Trimmomatic: A flexible read trimming tool for Illumina NGS data, (n.d.). http://www.usadellab.org/cms/?page=trimmomatic (accessed August 3, 2022).

[5] A. Dobin, C.A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, T.R. Gingeras, STAR: ultrafast universal RNA-seq aligner, Bioinforma. Oxf. Engl. 29 (2013) 15–21, doi:10.1093/bioinformatics/bts635.

[6] S. Forrester, A. Goundry, B.T. Dias, T. Leal-Calvo, M.O. Moraes, P.M. Kaye, J.C. Mottram, A.P.C.A. Lima, Tissue Specific Dual RNA-Seq Defines Host-Parasite Interplay in Murine Visceral Leishmaniasis Caused by Leishmania donovani and Leishmania infantum, Microbiol. Spectr. 10 (2022) e0067922, doi:10.1128/spectrum.00679-22.

[7] M. Aslett, C. Aurrecoechea, M. Berriman, J. Brestelli, B.P. Brunk, M. Carrington, D.P. Depledge, S. Fischer, B. Gajria, X. Gao, M.J. Gardner, A. Gingle, G. Grant, O.S. Harb, M. Heiges, C. Hertz-Fowler, R. Houston, F. Innamorato, J. Iodice, J.C. Kissinger, E. Kraemer, W. Li, F.J. Logan, J.A. Miller, S. Mitra, P.J. Myler, V. Nayak, C. Pennington, I. Phan, D.F. Pinney, G. Ramasamy, M.B. Rogers, D.S. Roos, C. Ross, D. Sivam, D.F. Smith, G. Srinivasamoorthy, C.J. Stoeckert Jr, S. Subramanian, R. Thibodeau, A. Tivey, C. Treatman, G. Velarde, H. Wang, TriTrypDB: a functional genomic resource for the Trypanosomatidae, Nucleic Acids Res. 38 (2010) D457–D462, doi:10.1093/nar/gkp851.