Online Supplementary Material for:

Blood transcriptomes reveal novel parasitic zoonoses circulating in Madagascar's lemurs

Peter A. Larsen, Corinne E. Hayes, Cathy V. Williams, Randall E. Junge, Josia Razafindramanana, Vanessa Mass, Hajanirina Rakotondrainibe, and Anne D. Yoder

S1. Molecular Methods

Whole blood from three *Indri indri* and three *Propithecus diadema* was preserved in Tempus™ Blood RNA Tubes (Life Technologies, Grand Island, NY). Total RNA was extracted from each sample using the Tempus™ Spin RNA Isolation Kit (Life Technologies, Grand Island, NY). Globin mRNA was reduced using the GLOBINclear Kit (Life Technologies, Grand Island, NY) and custom designed lemur globin capture oligo mix (10 pmol/ul; Table S3). Globin depleted RNA was then purified using the RNA Clean & Concentrator Kit (Zymo Research, Irvine, CA). RNA quality was checked using an Agilent 2100 Bioanalyzer and RNA integrity numbers ranged from 7.3 to 9.3 (Figure S1). For each sample, 100ng of RNA was used for RNA-seq library construction (~300 bp fragment size) following the Illumina Stranded RNA-Seq + Ribozero Gold (Human/Mouse/Rat) protocol. Samples were barcoded, pooled, and sequenced on one Illumina HiSeq 2000 lane (100bp PE). Illumina library preparation and sequencing was performed at the Duke Genome Sequencing Shared Resource (Duke Center for Genomic and Computational Biology, Duke University). All raw data generated for this study have been deposited in the Sequence Read Archive under BioProject number PRJNA293089.

S2. Bioinformatics

Raw reads were quality filtered (including Illumina adapter removal) using Trimmomatic v0.32 software [1]. Leading and trailing bases with quality scores < 20 were trimmed from each read and a sliding window of 4 bases with minimum 20 was implemented (minimum trimmed sequence length 70bp). Illumina HiSeq sequencing resulted in approximately 37 million 100bp read pairs per sample and, of these, ~76% survived quality filtering (Table S4).

For each individual, BBMap software (www.bbmap.sourceforge.net) was used to map filtered reads to the *Microcebus murinus* draft genome (GenBank assembly accession: GCA_000165445.1) and all unmapped reads were retained for downstream analyses (Table S4). Using this approach, approximately 32% of all reads mapped and an average of 38.25 million unmapped reads were used for *de novo* transcriptome assembly for each of the six samples (Table S4). *De novo* transcriptome assemblies of unmapped reads were performed using the Trinity v2.0 software package [2]. Approximately 352,000 assembled transcripts were observed across the six samples (Table S5). Each transcriptome assembly was imported into the Galaxy software platform [3-5] and was filtered to a minimum contig length of 300 bp. Transcriptome contigs were then grouped according to taxonomic classification within Galaxy using Megablast for preliminary taxonomic identifications (nt database, word size = 16, perc_identity = 95.0, evalue = 1e-06). All non-mammalian reads were screened for vector-borne parasites and associated sequences were retained for downstream phylogenetic analyses. The Megablast results provided an initial taxonomic identification to the family and genus level. Using this workflow, we identified *de novo* assemblies of ribosomal RNA and mitochondrial genes for parasites discussed herein.

The Trinity assemblies were then independently confirmed by mapping the filtered Illumina sequence data to closely related genomes/sequences (identified using the Megablast output described above) with the TopHat software package (https://ccb.jhu.edu/software/tophat/index.shtml; Table S6) [6]. TopHat mapping results are presented in Table S6. For each of the six individual blood transcriptomes, reads mapping to ribosomal 18S and 28S (*Babesia*) 16S and 23S (*Borrelia* and *Candidatus* Neoehrlichia), 18S and 28S-Alpha (*Trypanosoma*) and the mitochondrial cytochrome oxidase I and cytochrome-*b* genes (*Plasmodium*) were aligned to Trinity contigs. This process allowed us to 1) confirm the accuracy of the *de novo* Trinity contigs 2) examine intra-individual genetic variation of the pathogens that might indicate multiple strains circulating within a single host. Maximum Likelihood phylogenetic analyses were performed using available rRNA genes and the mitochondrial COI gene for closely related sequences available in the published literature and on GenBank (Fig. 2; FigS2) using RAxML v.8 software (-m GTRGAMMAI -# 500) [7]. Transcriptome contigs used for phylogenetic analyses were submitted to GenBank under the accession numbers (KT722781-KT722795) and all contigs assembled using filtered RNA-Seq data are provided on Dryad. [Number pending acceptance]

S3. Supplementary References

1 Bolger, A.M., Lohse, M. & Usadel, B. 2014 Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, btu170.

2 Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R. & Zeng, Q. 2011 Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature biotechnology* **29**, 644-652.

3 Goecks, J., Nekrutenko, A. & Taylor, J. 2010 Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol* **11**, R86.

4 Blankenberg, D., Kuster, G.V., Coraor, N., Ananda, G., Lazarus, R., Mangan, M., Nekrutenko, A. & Taylor, J. 2010 Galaxy: a web - based genome analysis tool for experimentalists. *Current protocols in molecular biology*, 19.10. 11-19.10. 21.

5 Giardine, B., Riemer, C., Hardison, R.C., Burhans, R., Elnitski, L., Shah, P., Zhang, Y., Blankenberg, D., Albert, I. & Taylor, J. 2005 Galaxy: a platform for interactive large-scale genome analysis. *Genome research* **15**, 1451-1455.

6 Trapnell, C., Pachter, L. & Salzberg, S.L. 2009 TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 1105-1111.

7 Stamatakis, A. 2014 RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312-1313.

8 Lack, J.B., Reichard, M.V., and Van Den Bussche, R.A. 2012. Phylogeny and evolution of the Piroplasmida as inferred from 18S rRNA sequences. *International Journal of Parasitology* **42**, 353-363.

9 Lima, L., Espinosa-Álvarez, O., Hamilton, P.B., Neves, L., Takata, C.S., Campaner, M., Attias, M., de Souza, W., Camargo, E.P. & Teixeira, M.M. 2013 Trypanosoma livingstonei: a new species from African bats supports the bat seeding hypothesis for the Trypanosoma cruzi clade. *Parasit Vectors* **6**, 221.

S4. Supplementary Figures

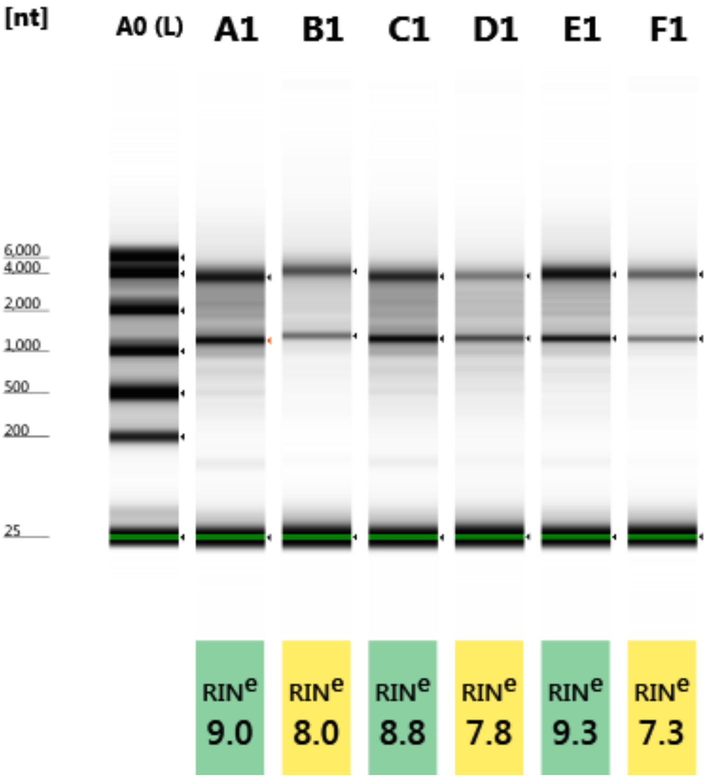

Figure S1. Agilent Bioanalyzer quality RIN scores for six lemur RNA blood extractions. Lane identifications (see Fig 1 and Table 1 in main text): A1: *Indri indri* 1; B1: *Propithecus diadema* 1, C1: *I. indri* 2; D1: *I. indri* 3; E1: *P. diadema* 2; F1: *P. diadema* 3.

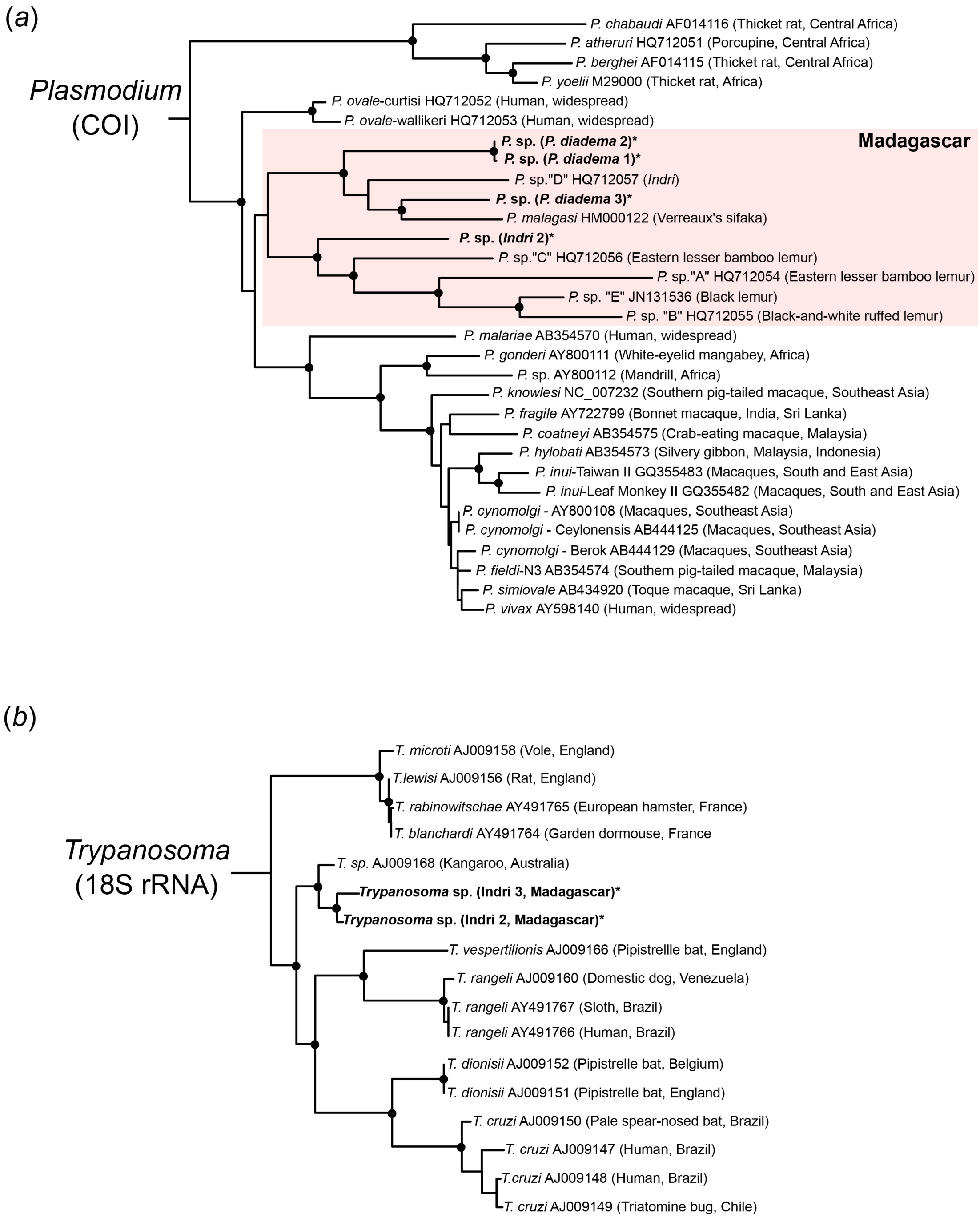
(a)
Plasmodium (COI)
P. chabaudi AF014116 (Thicket rat, Central Africa)
P. atheruri HQ712051 (Porcupine, Central Africa)
P. berghei AF014115 (Thicket rat, Central Africa)
P. yoelii M29000 (Thicket rat, Africa)
P. ovale-curtisi HQ712052 (Human, widespread)
P. ovale-wallikeri HQ712053 (Human, widespread)
Madagascar
P. sp. (P. diadema 2)*
P. sp. (P. diadema 1)*
P. sp."D" HQ712057 (Indri)
P. sp. (P. diadema 3)*
P. malagasi HM000122 (Verreaux's sifaka)
P. sp. (Indri 2)*
P. sp."C" HQ712056 (Eastern lesser bamboo lemur)
P. sp."A" HQ712054 (Eastern lesser bamboo lemur)
P. sp. "E" JN131536 (Black lemur)
P. sp. "B" HQ712055 (Black-and-white ruffed lemur)
P. malariae AB354570 (Human, widespread)
P. gonderi AY800111 (White-eyelid mangabey, Africa)
P. sp. AY800112 (Mandrill, Africa)
P. knowlesi NC_007232 (Southern pig-tailed macaque, Southeast Asia)
P. fragile AY722799 (Bonnet macaque, India, Sri Lanka)
P. coatneyi AB354575 (Crab-eating macaque, Malaysia)
P. hylobati AB354573 (Silvery gibbon, Malaysia, Indonesia)
P. inui-Taiwan II GQ355483 (Macaques, South and East Asia)
P. inui-Leaf Monkey II GQ355482 (Macaques, South and East Asia)
P. cynomolgi - AY800108 (Macaques, Southeast Asia)
P. cynomolgi - Ceylonensis AB444125 (Macaques, Southeast Asia)
P. cynomolgi - Berok AB444129 (Macaques, Southeast Asia)
P. fieldi-N3 AB354574 (Southern pig-tailed macaque, Malaysia)
P. simiovale AB434920 (Toque macaque, Sri Lanka)
P. vivax AY598140 (Human, widespread)
(b)
Trypanosoma (18S rRNA)
T. microti AJ009158 (Vole, England)
T.lewisi AJ009156 (Rat, England)
T. rabinowitschae AY491765 (European hamster, France)
T. blanchardi AY491764 (Garden dormouse, France)
T. sp. AJ009168 (Kangaroo, Australia)
Trypanosoma sp. (Indri 3, Madagascar)*
Trypanosoma sp. (Indri 2, Madagascar)*
T. vespertilionis AJ009166 (Pipistrellle bat, England)
T. rangeli AJ009160 (Domestic dog, Venezuela)
T. rangeli AY491767 (Sloth, Brazil)
T. rangeli AY491766 (Human, Brazil)
T. dionisii AJ009152 (Pipistrelle bat, Belgium)
T. dionisii AJ009151 (Pipistrelle bat, England)
T. cruzi AJ009150 (Pale spear-nosed bat, Brazil)
T. cruzi AJ009147 (Human, Brazil)
T.cruzi AJ009148 (Human, Brazil)
T. cruzi AJ009149 (Triatomine bug, Chile)

Figure S2. ML phylogenies of *Plasmodium* (A) and *Trypanosoma* (B). Taxa in bold were identified from *I. indri* and *P. diadema* (table 1; figure 1). Black circles identify statistically supported nodes (>75% BS). *a*: Phylogeny based on mitochondrial COI gene sequence

variation. Red shading highlights a clade of *Plasmodium* endemic to Madagascar. *b*: Phylogeny based on 18S ribosomal RNA. The trypanosome sequences identified in our sample are within the '*T. cruzi* Clade' (sensu Lima *et al.* 2013 [9]) and share ~1.7% genetic similarity to an unidentified species from Australia.

Table S1. Number of base substitutions per site from averaging over sequence pairs between putative species of *Babesia* (see Figure 1A). Intra- (diagonal) and inter- (below diagonal) genetic distance values were generated using the Kimura 2-parameter model.

	1	2	3	4	5
1) *Babesia* sp. Madagascar	-				
2) *B.* sp. South Africa	0.020	0.007			
3) *B. leo*	0.021	0.010	0.006		
4) *B. rodhaini*	0.030	0.023	0.028	0.005	
5) *B. microti*	0.032	0.019	0.023	0.024	0.001

Table S2. Number of base substitutions per site from averaging over sequence pairs between putative species of *Borrelia* (see Figure 1A). Intra- (diagonal) and inter- (below diagonal) genetic distance values were generated using the Kimura 2-parameter model.

	1	2	3	4	5
1) *Borrelia* sp. Madagascar	-				
2) *B. theileri*	0.010	0.005			
3) *B. lonestari*	0.016	0.011	0.002		
4) *B. miyamotoi*	0.020	0.014	0.019	0.006	
5) *B. coriaceae*	0.021	0.016	0.018	0.020	0.004

Table S3. Lemur specific biotinylated probes used for hemoglobin reduction. Probes were designed based on predicted Alpha and Beta hemoglobin genes annotated in the genomes of *Microcebus murinus* (NCBI Genome ID 777) and *Propithecus coquereli* (NCBI Genome ID 24390).

Probe Name	**Sequence**
Mmurinus_Alpha1	/5Biosg/CTC CAG GGC CTC SGC GCC ATY GTC
Mmurinus_Alpha2	/5Biosg/TGG TGG TGG GGA AGG AMW GGA ACA
Mmurinus_Alpha3	/5Biosg/TGT CGA AGT GGG AGA AGT AGG TCT
Mmurinus_Beta1	/5Biosg/CAT AAY AGC AGA AGG AGA GGA CAG G
Mmurinus_Beta2	/5Biosg/CCA CAG AGA GST GAC ATG ASC A
PcoqAlpha1	/5Biosg/CTR CCA CCC ACT CAG ACT TTA TTC AA
PcoqAlpha2	/5Biosg/CCC AGT GCG TCG GCC MCC TTC TT
PcoqAlpha3	/5Biosg/GGT CGA AGT GGG GGA AGT AGG TCT
PcoqBeta1	/5Biosg/CCA CAG GCY GGT GAC CTG AGC A
PcoqBeta2	/5Biosg/CAT GAC AGC AGA AGG AGA GGA CAG G
PcoqBeta3	/5Biosg/CCA TCG CTA AAA GCA CTC AGC ACC

Table S4. Quality filtering of Illumina paired end sequences and BBEdit mapping results.

	Raw Read Pairs	Surviving Read Pairs	Mapped Reads	Unmapped Reads
I. indri 1	40250092	30691320	17047480	44335160
I. indri 2	41210410	31560150	20510476	42609824
I. indri 3	40547995	30228397	17610902	42845892
P. diadema 1	37966647	28809033	19334938	38283128
P. diadema 2	32356236	24067307	16468370	31666244
P. diadema 3	31620419	23940419	18117468	29763370

Table S5. Trinity *de novo* transcriptome assemblies of six blood samples. The Trinity software package groups *de novo* assembled transcripts into clusters (or 'genes') based on sequence similarity (Assembled Clusters column) and then identifies putative transcript variants (or isoforms) within those clusters (Assembled Cluster Transcripts column). The contig N50 statistic indicates that 50% of the assembled transcripts were at least 755 base pairs in length (averaged across the six assemblies).

	Total Assembled Bases	**Assembled Clusters**	**Assembled Cluster Transcripts**	**Contig N50 (bp)**
I. indri 1	211038196	320386	346321	769
I. indri 2	195066403	297049	321021	768
I. indri 3	236624006	356483	385263	779
P. diadema 1	237565028	357108	384818	790
P. diadema 2	209034719	328019	351557	737
P. diadema 3	188197365	306565	327746	692

Table S6. TopHat mapping results. Values in each column indicate the number of filtered Illumina RNA-Seq sequences that mapped to reference sequences (column 2).

Reference Species	**NCBI ID**	***I. indri* 1**	***I. indri* 2**	***I. indri* 3**	***P. diadema* 1**	***P. diadema* 2**	***P. diadema* 3**
Babesia microti	GCA_00691945	7,611	63,379	56,906	46,228	32,337	100,568
Borrelia miyamotoi	GCA_000445425	12,352	0	1	5	3	6
Ehrlichia ruminantium	GCA_000026005	208	26	1	2	2	3782
Trypanosoma cruzi	GCA_000209065	6,761	75,279	58,091	28,364	17,661	35,494
Plasmodium knowlesi	GCA_000006355	5,002	54,938	24,971	40,240	26,600	46,739

Table S7. Number of sequences mapping to ribosomal and mitochondrial genes for parasites discussed herein. The first value is the number of contigs, followed by number of supporting RNA-Seq sequences in parentheses, and finally the total length of contigs in base pairs.

	I. indri **1**	***I. indri*** **2**	***I. indri*** **3**	***P. diadema*** **1**	***P. diadema*** **2**	***P. diadema*** **3**
Babesia						
18S	5 (88) 609 bp	6 (2,244) 1,238 bp	6 (3,286) 1,233bp	5 (530) 1,040 bp	6 (932) 1,201 bp	5 (8,565) 1,491 bp
28S	7 (3,589) 905 bp	8 (13,968) 2,012 bp	13 (12,400) 2,158 bp	10 (7,954) 1,683 bp	8 (8,930) 1,880 bp	10 (17,052) 2,149 bp
Borrelia						
16S	2 (3,521) 1648 bp					
23S	5 (8,639) 2,828 bp					
C. **Neoehrlichia**						
16S	4 (35) 497 bp					7 (538) 900 bp
23S	7 (173) 899 bp					8 (3241) 1,740 bp
Plasmodium						
COI		1 (498) 1,431 bp		4 (227) 1,431 bp	1 (215) 1,431 bp	1 (311) 1,431 bp
Cyt-b		1 (496) 1,131 bp		2 (152) 1,131 bp	1 (127) 1,131 bp	1 (149) 1,131 bp
Trypanosoma						
18S		1 (348) 1,128 bp	4 (1,145) 1,140 bp			
28S-Alpha		3 (3,468) 1,421 bp	5 (3,129) 1,544 bp			