Comparative and systems analyses of *Leishmania* spp. non-coding RNAs through developmental stages

J. Eduardo Martinez-Hernandez ^{1,2,3}, Victor Aliaga-Tobar ^{3,4}, Carolina González ^{2,5}, Rubens Monte-Neto ⁶, Alberto J. M. Martin ^{2,7*}, Vinicius Maracaja-Coutinho ^{3,4,8*}

¹ Programa de Doctorado en Genómica Integrativa, Vicerrectoría de Investigación, Universidad Mayor, Santiago, Chile.

² Laboratorio de Biología de Redes, Centro de Genómica y Bioinformática, Facultad de Ciencias, Universidad Mayor, Santiago, Chile.

³ Centro de Modelamiento Molecular, Biofísica y Bioinformática – CM²B², Facultad de Ciencias Químicas y Farmacéuticas, Universidad de Chile, Santiago, Chile.

⁴ Advanced Center for Chronic Diseases - ACCDiS, Facultad de Ciencias Químicas y Farmacéuticas, Universidad de Chile, Santiago, Chile.

⁵ Center for Bioinformatics and Genome Biology, Fundación Ciencia & Vida, Santiago, Chile.

⁶ Instituto René Rachou - Fundação Oswaldo Cruz, Belo Horizonte, Minas Gerais, Brazil.

⁷ Escuela de Biotecnología, Facultad de Ciencias, Universidad Mayor, Chile.

⁸ Instituto Vandique, João Pessoa, Brazil.

* Corresponding authors.

E-mail: alberto.martin@umayor.cl (AJMM)

E-mail: vinicius.maracaja@uchile.cl (VMC)

28 **ABSTRACT**

29 *Leishmania* spp. is the causal agent of several diseases called leishmaniasis, neglected
30 diseases that seek to be eradicated in the coming years. We aimed to study the genomic
31 structure and function of non-coding RNAs (ncRNAs) from *Leishmania* spp. and to get
32 insights into its RNAome. We studied 26 strains corresponding to 16 different species of
33 *Leishmania* genus. RNAome analysis revealed the presence of several ncRNAs that are
34 shared through different species, allowing us to differentiate between subgenus and as well
35 as species that are canonically related to visceral leishmaniasis. We found coexpression
36 relationships within coding genes and ncRNAs, thus suggesting possible functional
37 relationships between different coding genes-ncRNAs. Expression analysis in the metacyclic
38 developmental stage comparison for *Leishmania braziliensis* and *Leishmania major* reveals
39 the presence of shared coexpressed ncRNAs to several other coding genes in both species
40 involved in chromatin structure and host interaction. This work constitutes the first effort to
41 characterize the *Leishmania* RNAome, supporting further approaches to better understand
42 the role of ncRNAs in the gene regulation, infective process and host-parasite interaction.

43

44 **AUTHOR SUMMARY**

45 Leishmaniasis is a worldwide tropical neglected disease, which is endemic in several
46 countries of America, Africa, Asia and Europe. *Leishmania* parasites are transmitted to
47 mammal hosts by the bite of female sandflies. Currently, more than one million of the new
48 infections are reported mainly in the Americas, the Mediterranean basin, the Middle East and
49 Central Asia. The genomic knowledge of these parasites have increased since the first draft
50 genome was published in 2005, however the role of non-coding RNAs in pathogenesis and
51 development is still an open avenue. Here we use a system and comparative genomic
52 approach to characterize and predict the function of the whole ncRNA repertoire in
53 *Leishmania* parasites. Although *Leishmania* presents a large amount of ncRNA, there is a

low conservation rate, however it is directly proportional to phylogenetic relationships. We also found that this conserved ncRNA was related to histone regulation and proteophosphoglycan ppg3 co-expression. Our study provides new insights in our knowledge about function and expression patterns of ncRNA and their relationship with *Leishmania* chromatin regulation and virulence.

INTRODUCTION

Leishmaniasis is a group of different diseases caused by *Leishmania* spp. parasites (1–3). These are classified based on their wide clinical manifestations, that range from self-cured skin lesions (cutaneous leishmaniasis or CL) to the visceral form (visceral leishmaniasis or VL) affecting organs such as liver, spleen and bone marrow (1,4). Reported annual infections vary from 0.7 to 1.3 million for CL and 300,000 cases for VL (5).

The life cycle of *Leishmania* parasites involves invertebrate and mammal hosts, with several morphological changes and biochemical adaptations (6). Mammalian infection by *Leishmania* begins with the entry of metacyclic promastigotes (META) into the dermis by the bite of a sandfly, a blood-feeding Diptera part of the Phlebotominae clade. Metacyclic promastigotes are rapidly internalized by skin resident or recruited phagocytes notably neutrophils, macrophages and dendritic cells (7). Within the host cells *Leishmania* parasites trigger the maturation and morphological transformation to amastigote (AMA) forms. At this point the parasites multiply by binary fission. This division leads to the lysis of the macrophage and induces the dissemination of amastigotes, allowing them to infect other cells (7–9). Finally, parasites re-enter into the vector by blood meal of an infected host, taking with them macrophages carrying amastigotes, triggering morphological changes to procyclic promastigotes (PRO) (10). Such a challenging life cycle requires parasite adaptation efforts to survive under nutritional starvation, sudden changes in temperature, pH and cellular plasticity varying from free living 15-30 µm length flagellated form to a 3-6 µm

80 rounded intracellular form. All these can be achieved by a very well orchestrated
81 post-transcriptionally controlled gene expression and epigenetic events as regulatory
82 mechanisms already reported in trypanosomatids (11,12).

83 Non-coding RNAs (ncRNAs) are untranslated transcripts that participate modulating multiple
84 biological processes, such as gene regulation at transcriptional (13) and post-transcriptional
85 levels (14,15), developmental processes (16) or even in many diseases, such cancer (17) or
86 pathogenic infections (18). ncRNAs can be classified into two main subclasses (17,19,20).
87 The small RNAs (sncRNAs) (< 200 nt), a subclass that includes regulatory RNAs such
88 microRNAs (miRNAs), small interfering RNAs (siRNAs), Piwi-interacting RNA (piRNAs),
89 while the long non-coding RNAs (lncRNAs) (> 200 nt) includes several transcripts known to
90 take place regulating several important processes in eukaryotes, such as genome imprinting
91 (21), splicing (22) and chromatin organization (23). Many classes of regulatory ncRNAs have
92 been virtually described in all eukaryotic cells since the miRNAs were first described in 1993
93 in *Caenorhabditis elegans* (24). Although ncRNAs have been extensively studied in recent
94 years, little is known about their roles in the development and pathogenesis of eukaryotic
95 microbes (25).

96 Since the first genome of *Leishmania major* was sequenced in 2005 (26), many studies
97 addressed the understanding of the genome structure and function at protein coding genes
98 level (26–35). In parallel, studies focused on ncRNAs in *Leishmania* parasites have
99 increased since its first description in 2006 (36). These works have been mainly
100 concentrated on the identification of specific ncRNAs classes, such as siRNA (37);
101 microRNA-like and their regulatory roles (38); small RNAs derived from tRNAs and rRNAs
102 as regulators of host-pathogen interaction processes (39); small nucleolar RNAs (snoRNA)
103 and their function in rRNA processing (40); UTR-associated ncRNAs (uaRNAs) (41); or
104 identifying lncRNAs and their putative functions (42). Recently, two works devoted to
105 characterizing the ncRNAs repertoire in *L. braziliensis* brought new challenges on the

106 understanding of non-coding transcripts' role in trypanosomatids developmental life cycle
107 stages (43,44). Certain lncRNAs act in *Trypanosoma brucei* differentiation and survival within
108 the host (45). In this regard, a system biology approach has not yet been performed in
109 *Leishmania* parasites.

110 In this work, we combined different computational, transcriptome, comparative genomics and
111 systems biology approaches in order to identify and characterize the ncRNAs repertoire in
112 26 different strains belonging to 16 *Leishmania* species. We focused on the identification of
113 regulatory ncRNAs that are shared or unique within different *Leishmania* species. We also
114 analyzed the expression of these ncRNAs and their relationship with coding genes, in order
115 to identify associated functions through coding/non-coding co-expression networks.
116 Interestingly, we observed several ncRNA-coding RNA pairs co-expressed in the same
117 developmental stages. These pairs were differentially expressed (DE) in consecutive
118 developmental steps, revealing the preponderant role of ncRNAs in the regulation of
119 development and survival of *Leishmania* parasites.

120

121 RESULTS

122 Dissecting the repertoire of non-coding RNAs across *Leishmania* spp.

123 We performed a computational experiment to identify the ncRNA repertoire of all evaluated
124 strains, by mapping known ncRNA sequences available in public databases and searching
125 established covariance models against 26 *Leishmania* genomes. Our approach
126 characterized, for the first time, a plethora of different types of ncRNA in 16 species of
127 *Leishmania* genus. Our ncRNA predictions revealed a total of 105,695 ncRNAs, ranging
128 from 2,424 to 6,213 per species, in which *Leishmania turanica* LEM423 and *Leishmania* sp.
129 LD974 presented, respectively, the higher and lower number of predicted ncRNAs
130 (Supplementary Table 1). We identified 19 well known RNA classes spreaded over
131 *Leishmania* spp. genomes, according to NRDR/NR2 (46) and Rfam 14 (47) original

132 annotations and nomenclature; and an additional set of ncRNAs without any specific
133 functional annotation but conserved according to other RNA sequences available in public
134 databases and indexed in NR2. The most representative RNA class is cis-Regulatory RNA,
135 with an average of 1,122 copies per genome. The less abundant class is scaRNA and
136 ta-siRNA, presenting only 1 copy in LspLD974 and LtarParTarII, respectively. The top ten
137 RNA classes according to the copy number identified in average/genomes were:
138 cis-Regulatory RNA (1,122), piRNA (862), rRNA (790), miRNA (559), Unclassified (290),
139 siRNA (174), tRNA (139), H/ACA box snoRNAs (63), snoRNA (34) and C/D box snoRNAs
140 (9). As observed, miRNAs, piRNAs and cis-reg RNAs correspond to more than 50% of the
141 ncRNAs found in each evaluated genome (Figure 1A, Supplementary Table 3). It is worth
142 noting that the ncRNAs presented a total GC content ranging within 48% and 60%, and a
143 length ranging from 18 (miRNA) to 6,212 nt (rRNA).

144 In order to evaluate the transcriptional evidence for predicted ncRNAs, we used publicly
145 available RNA-seq data (Supplementary Table 2) of four strains from three different species:
146 *L. braziliensis* M2903 (MHOM/BR/75/M2903) and *L. braziliensis* M2904
147 (MHOM/BR/75/M2904), *L. donovani* BPK282A1 (MHOM/NP/02BPK282) and *L. major*
148 Friedlin (MHOM/IL/81/Friedlin). After applying an expression cut-off of 0.5 CPM to be
149 considered as a bona fide expressed transcript, we were able to obtain transcriptional
150 evidence for the majority of the predicted ncRNAs. Our analysis revealed that the
151 percentage of expressed ncRNAs range between 83% for *L. donovani* to 95% in *L.*
152 *braziliensis* (Figure 1B, Supplementary Table 4). Further details of ncRNAs repertoire for
153 each strain are described in Supplementary Table 1 and Supplementary Table 3.

154 After filtering all normalized values of the top representative ncRNAs, we compared their
155 expression values ($\log_2$ Counts Per Million) within them and with the protein coding genes
156 for each strain (Figure 1C). As expected, we identified a mean value of CPM for ncRNAs
157 ranging from 4.95 (snoRNA) to 11.65 (C/D box snoRNA) in *L. braziliensis* (Supplementary

Table 5), corresponding an average between 464 and 593 read counts, respectively. On the contrary, we observed a mean CPM value of 107.87 for coding genes, corresponding to an average of 18,781 reads per gene. As observed in other species, the mean expression value for coding genes are in general higher than for ncRNA classes. The only exceptions are siRNA, miRNA, miRNA cis-regulatory and unclassified ncRNAs for *L. major* and cis-regulatory ncRNA for *L. donovani*. In Figure 1C and Supplementary Table 5 we describe all other results for this comparison.

165

The *Leishmania* spp. pan-RNAome: ncRNAs conservation across 26 genomes.

We carried out a comparative genomic analysis in order to find patterns of conservation throughout the ncRNA repertoire in all *Leishmania* strains. The complete inventory of 105,695 ncRNAs was grouped into 12,012 clusters, which were named as the pan-RNAome of *Leishmania* spp. The pan-RNAome was divided into a core-RNAome of only 134 (1.1%) clusters, composed by 23,640 ncRNAs shared among all species; and a total of 11,878 (98.9%) clusters of 77,174 accessory ncRNAs, shared within 2 and 25 genomes (Figure 2, Supplementary Figure 1). tRNA, rRNA, cis-regulatory RNA and miRNA were the major classes of ncRNAs that pertain to the core-RNAome (Supplementary Figure 1). Curiously, 9% of the core-ncRNAs did not present a functional annotation and were annotated as "Unclassified ncRNAs". These were mainly ncRNA transcripts deposited in public databases as RNAs with a yet to be characterized function.

From the accessory RNAome, we identified a set of 4,881 ncRNAs (4,536 clusters) that are unique for each particular evaluated *Leishmania* isolate, representing 38% of the pan-RNAome. This subset of unique ncRNAs is the most diverse in terms of ncRNA types. Nonetheless, the most common classes were cis-regulatory RNAs, miRNAs and piRNAs, which together represent 71% of the total inventory of this subset (Supplementary Figure 1). Interestingly, we identified a subset of accessory ncRNAs that allowed us to distinguish

184 between *Viannia* and *Leishmania* subgenus. The *Viannia* subgenus species shared 224
 185 clusters of ncRNAs that are absent in the other groups (Figure 2). The *Viannia*
 186 subgenus-specific clusters are mainly annotated as gene expression regulators ncRNAs
 187 (Supplementary Figure 1), with 66% of them belonging to either cis-regulatory RNAs,
 188 miRNAs, piRNAs or siRNAs. In contrast, *Leishmania* (*Leishmania*) subgenus shared only 3
 189 ncRNA clusters, with a piRNA and two rRNAs shared within all species of this group. Next,
 190 we evaluated the ncRNAs conservation in species canonically related to VL. For that, we
 191 reviewed several works and selected *Leishmania donovani* and *Leishmania infantum* as
 192 representative of VL classically associated species (2,48,49). Based on this, we found a
 193 subset of 111 ncRNA clusters (Figure 2), which belongs to cis-regulatory RNAs, rRNAs,
 194 miRNAs and piRNAs (Supplementary Figure 1).

195 To evaluate the relationship between the phylogeny based on protein coding genes and
 196 ncRNA classes distribution within all *Leishmania* species that cause different clinical
 197 manifestations or subgenus, we compared the phylogenetic relationship based on the
 198 sequences of 613 core orthologous genes conserved within all 26 *Leishmania* spp. isolates,
 199 with the hierarchical clustering generated based on the ncRNA presence/absence matrix.
 200 Interestingly both approaches, sequence phylogeny and binary matrix clustering, showed a
 201 similar behavior, demonstrating a phylogenetic relation in the conservation and distribution of
 202 ncRNAs (Supplementary Figure 2).

203

204 **Co-expression network analysis identifies developmental stages-associated gene** 205 **modules in *Leishmania* spp.**

206 We combined expression matrices of coding genes and ncRNAs processed in an analogous
 207 way from the same samples to perform weighted gene co-expression network analysis
 208 (WGCNA) (50). rRNA and tRNA classes were excluded due to their ubiquitous presence and
 209 expression patterns. The constructed network for *L. braziliensis* M2903 strain was obtained

210 using a total of 12,516 genes (sum of coding genes and ncRNAs) and 36 RNA-seq samples,
 211 corresponding to procyclic promastigote (PRO), metacyclic promastigote (META) and
 212 amastigote (AMA). For *L. donovani* were included 10,738 genes and 7 RNA-seq samples,
 213 obtained from PRO and AMA. Likewise, we used 12,506 genes and 11 RNA-seq samples
 214 corresponding to META and PRO in order to construct a co-expression network for *L. major*.
 215 In all three analyses we used a scale-free topological index of 0.85. The soft threshold
 216 values were setted to 30, 26 and 12 respectively for *L. braziliensis*, *L. donovani* and *L. major*.
 217 Different numbers of co-expressed modules were identified for all three species, shown in
 218 Figure 3. Our results showed a different number of modules for each species. For *L.*
 219 *braziliensis* we identified 10 modules of which modules M9, M3 and M4 presented the higher
 220 number of protein coding genes, besides M10, M9 and M3 are the top 3 modules with the
 221 biggest number of ncRNAs. On the other hand, a total of 19 and 18 modules were obtained
 222 for *L. donovani* and *L. major*, respectively; and the module M18 presented the higher
 223 number of coding genes and ncRNAs in *L. donovani*. In Contrast, M15 of *L. major* were
 224 composed for the bigger number of coding genes (3,702) and ncRNAs (1,696) genes. The
 225 numbers of coding genes and ncRNAs present in each module were represented on Table 1.
 226 To analyze the correlation of each module and different *Leishmania* spp. developmental
 227 stages, we used a module-development stage relationship comparison. In Figure 3, we
 228 present all module-developmental stage relationship results. Finally, we identified the genes
 229 (coding and ncRNAs) that were highly correlated to different developmental stages in
 230 *Leishmania* parasites by applying gene significance (GS) and module membership (MM)
 231 measures. In Table 2, we describe the total number of genes highly correlated to AMA, PRO
 232 and META for each one of the three species and related enriched modules. These results
 233 provide evidence that the development of *Leishmania* parasites requires the joint expression
 234 of different types of coding and non-coding genes.

235

236

Table 1. Number of total coding genes and ncRNAs per module in different *Leishmania* parasites

Module	<i>L. braziliensis</i>		<i>L. donovani</i>		<i>L. major</i>	
	coding	ncRNA	coding	ncRNA	coding	ncRNA
M1	163	45	67	21	548	205
M2	6	44	43	21	1343	511
M3	1854	413	55	33	181	69
M4	1448	294	1913	304	195	61
M5	11	25	187	57	112	48
M6	5	54	89	32	159	83
M7	13	62	869	184	593	123
M8	11	26	94	42	78	91
M9	4847	606	52	132	259	61
M10	870	1523	428	90	204	60
M11	-	-	461	51	292	167
M12	-	-	285	364	40	63
M13	-	-	33	29	68	52
M14	-	-	64	37	297	108
M15	-	-	110	38	3702	1696
M16	-	-	941	226	53	120
M17	-	-	205	158	73	110
M18	-	-	1985	1320	21	42
M19	-	-	135	82	-	-

237

238

Table 2. Coding genes and ncRNAs presenting high correlation to different developmental stages in *Leishmania* parasites

species	AMA			META			PRO		
	coding	ncRNA	Module	coding	ncRNA	Module	coding	ncRNA	Module
<i>L. braziliensis</i>	2834	181	M3 M4 M9	889	44	M3 M4	1081	33	M1 M3 M4 M8 M9 M10

<i>L. donovani</i>	1693	24	M3 M4 M5 M8 M11 M15 M17 M18	-	-		1216	961	M1 M3 M4 M5 M7 M11 M15 M16
<i>L. major</i>	-	-		2012	1123	M1 M2 M3 M4 M5 M7 M9 M10 M11 M13 M14 M15 M16	2234	403	M1 M2 M3 M4 M5 M7 M M9 M10 M11 M14 M15 M16 M17

239

240 Functional associations between ncRNAs and coding genes

241 We performed GO terms enrichment analysis with Goseq (51) on the modules found in the
242 previous section that are significantly correlated in each species. After the filtering of
243 enriched networks, we made a functional enrichment analysis using GO terms for the linked
244 coding genes. The significantly enriched functional terms associated with ncRNAs for each
245 species are displayed in Figure 4, Supplementary Data 1, Supplementary Data 2 and
246 Supplementary Data 3. By following this procedure, we were able to assign different
247 numbers of ncRNAs to a variety of different functional categories. We assigned functions to
248 104 ncRNAs for *L. braziliensis* (Supplementary Data 1), 263 for *L. donovani* (Supplementary
249 Data 2) and 517 for *L. major* (Supplementary Data 3). When considering Biological Process
250 (BP) terms, ncRNAs in *L. braziliensis* (Figure 4A) were highly linked to genes enriched in
251 “biosynthetic process”, “nitrogen compound metabolic process”, “cellular metabolic process”
252 and “protein folding”. On the other hand, using the Molecular Function (MF) GO terms,
253 ncRNAs were linked to “protein binding” and “structural constituent of the ribosome”.
254 Likewise, the main Cellular Component (CC) associated GO terms were “cell part”,
255 “non-membrane-bounded organelle”, “organelle” and “ribonucleoprotein complex”. Alike *L.*

256 *donovani* ncRNAs (Figure 4B) were annotated in the main major BP GO terms associated
 257 with “cellular metabolic process”, “establishment of localization”. According to MF GO terms,
 258 these were mainly linked to “hydrolase activity” and “transmembrane transporter activity”, as
 259 well as “membrane part” and “proteasome complex” for major CC terms. For *L. major*
 260 ncRNAs (Figure 4C), the main BP GO annotations were “cellular metabolic process”,
 261 “biosynthetic process”, “macromolecule metabolic process” and “nitrogen compound
 262 metabolic process”. However, when considering MF GO annotations, these were linked to
 263 “transmembrane transporter activity”, “hydrolase activity” and “protein binding”. Finally,
 264 mainly CC GO terms were associated with “intracellular organelle”, “DNA packaging
 265 complex” and “protein-DNA complex”.

266

267 **Gene conservation through co-expression modules at different stages of parasitic** 268 **development reveals a potential association with genes related to host-pathogen** 269 **communication**

270 We also defined gene conservation through the three evaluated species and their
 271 corresponding modules for distinct developmental stages. For these, we compared modules
 272 of *L. braziliensis* and *L. donovani* and modules for the *L. braziliensis* and *L. major* species
 273 (Figure 5, Supplementary Data 4 and Supplementary Data 5). Interestingly, we found more
 274 coding genes conservation between AMA (1387 shared clusters) related modules, compared
 275 to META (783) or PRO (84) modules. On the other hand, we observed more conservation
 276 between META related ncRNAs (105), and less conservation in PRO related modules (7).
 277 However, we found a high number of no conservation regarding ncRNAs between
 278 developmental stages in *Leishmania* parasites.

279 We then applied GO terms enrichment analysis in order to obtain functional insights for the
 280 ncRNAs shared within clusters. We identified the potential function for 140 ncRNAs
 281 associated with different gene ontologies of previously mentioned genes (Supplementary

282 Data 6-8). For those clusters shared in AMA stage, we identified an enrichment for BP GO
 283 terms potentially related to host-pathogen communication among other functional categories
 284 mainly related to “symbiosis process”, “interaction between organisms”, “cellular component
 285 organization”, “cell communication”, “cell adhesion”, “regulation of biological process”,
 286 “regulation of cellular process”. Meanwhile, MF GO terms were enriched in “enzyme
 287 regulator activity”, “hydrolase activity”, “small molecule binding”, “ion binding” and
 288 “heterocyclic compound binding”. Regarding the CC GO terms, we obtained enrichment in
 289 “DNA packaging complex”, “host cellular component”, “non-membrane-bounded organelle”,
 290 “other organism part” and “protein-DNA complex” (Figure 5G). For the clusters shared in
 291 META stage (Figure 5H), we obtained the terms “cellular component organization”, “cellular
 292 process” and “organelle organization” for BP GO; with only “protein binding” enriched for MF.
 293 Meanwhile, CC GO terms enriched in META were “DNA packaging complex”,
 294 “non-membrane-bounded organelle” and “protein-DNA complex”. Likewise, for the clusters
 295 shared in PRO stage (Figure 5I), the BP GO terms enriched were “biosynthetic process”,
 296 “cellular metabolic process”, “macromolecule metabolic process”, “nitrogen compound
 297 metabolic process”, “organic substance metabolic process” and “primary metabolic process”.
 298 “structural constituents of the ribosome” was the unique term enriched for MF, meanwhile
 299 terms related to “cell part”, “intracellular organelle”, “non-membrane-bounded organelle”,
 300 “organelle” and “ribonucleoprotein complex” were the enriched CC terms.

301

302 **Differential expression of coding genes and non-coding RNAs through lifestyle stages** 303 **and their modules relationship**

304 To assess the variation in the expression patterns of ncRNAs and coding genes, and their
 305 relationship with previous findings across the three different developmental stages, we
 306 compared transcriptome experiments performed from PRO, META and AMA stages in *L.*
 307 *braziliensis* (PRO, META and AMA), *L. donovani* (AMA and PRO) and *L. major* (PRO and

308 META). For *L. braziliensis*, we analyzed the expression of 3,247 ncRNAs and 9,269, in
 309 pairwise comparison through AMA, META and PRO developmental stages. We identified 25
 310 ncRNAs overexpressed in AMA, compared to META. In this analysis, miRNAs (7),
 311 cis-regulatory RNAs (6) and piRNAs (5) represented the top classes. In opposite, for META
 312 we obtained 46 overexpressed ncRNAs, with cis-regulatory RNAs (24) appearing as the
 313 most DE class in this developmental stage, followed by piRNAs, Unclassified RNAs and
 314 miRNAs (Figure 6A and Supplementary Data 9). Considering coding genes, we observed 80
 315 overrepresented in AMA compared to 19 in META. When comparing the developmental
 316 stage AMA with PRO, we detected 156 and 316 ncRNAs and coding genes, respectively, as
 317 overexpressed in AMA, with cis-regulatory RNAs (46) presented as the more abundant
 318 class, followed by piRNAs (44) and miRNAs (29). In contrast, 17 ncRNAs and 38 coding
 319 genes were identified as overexpressed in PRO (Figure 6). Finally, by comparing the
 320 developmental stage META with PRO, we identified 116 ncRNAs overexpressed in META
 321 compared to only 4 overrepresented in PRO. Classes such as cis-regulatory RNAs (42) and
 322 piRNAs (30) were the majority represented in META. Likewise, we observed the expression
 323 of 130 coding genes overexpressed in META, compared to 9 in PRO.

324 In a similar approach, we evaluated the differential expression in *L. donovani*, analyzing
 325 2,714 ncRNAs and 8,024 coding genes but only comparing AMA and PRO stages. We used
 326 only these two developmental points due to the availability of RNA-seq data. We obtained an
 327 overexpression of 54 ncRNAs in AMA compared to 73 in PRO. In this comparison, we noted
 328 that cis-regulatory RNAs and HACA-box snoRNAs were the more expressed classes in
 329 AMA, meanwhile cis-regulatory RNAs and piRNAs were overrepresented in PRO. In a
 330 similar way, we detected the expression of 80 coding genes overexpressed in AMA,
 331 compared to 180 in PRO (Figure 6B and Supplementary Data 10).

332 We also evaluated the differential expression of ncRNAs and coding genes from *L. major*, by
 333 comparing PRO and META developmental stages. Our results showed a major expression

334 of ncRNAs in META developmental stage (Figure 6C and Supplementary Data 11), classes
335 such as cis-regulatory RNAs (115), miRNAs (51) and piRNAs (103) represented the 82% of
336 all ncRNAs differentially expressed in this stage. On the other hand, PRO stage presented
337 134 overrepresented ncRNAs, in which 97% is represented by the same three classes
338 overexpressed in META stage.

339 We additionally evaluated the appertain to modules of the differentially expressed genes and
340 ncRNAs. For instance, in *L. braziliensis* we observed that all differentially expressed ncRNAs
341 belong exclusively to two co-expression modules: M3 and M4 (Figure 7A). For AMA-META
342 comparison, 21 ncRNAs exclusively expressed in AMA were identified as part of M3 module,
343 of which miRNA class was the most represented class with 6 members overrepresented in
344 AMA, followed by cis-regulatory RNAs (4), piRNAs (3) and Unclassified RNAs (3). Contrary,
345 M3 module in PRO stage presented an overrepresentation of Cis-regulatory RNAs (17),
346 followed by piRNAs (4) and Unclassified RNAs (4), miRNAs (3) and siRNAs (3)
347 (Supplementary Data 9). This module corresponds to 72% (43) of all ncRNA expressed in
348 META (Figure 7). Likewise, we identified more module diversity in the same comparison for
349 coding genes, with the modules M9 and M10 containing the higher number of DE coding
350 genes. In this analysis, we obtained 33 DE coding genes for M9 in META-AMA comparison
351 (13 META and 20 AMA) and 40 in M10 (4 META and 36 AMA) (Figure 7A). A similar
352 behavior was observed for AMA-PRO and META-PRO comparisons. In AMA-PRO, the
353 module M9 represents 63% (24) of all overrepresented protein coding genes in the PRO
354 developmental stage, meanwhile in AMA stage it represents 44% (139). Moreover, the
355 module M10 presented only 5 genes from M10 in PRO stage, besides 113 genes in AMA
356 stage (Figure 7A). In META-PRO comparison, the module M9 with 5 coding genes
357 represents 55% of the total of overrepresented protein coding genes in the PRO stage,
358 besides the module M9 in META stage is equivalent to 42% (58) of total overexpressed
359 protein coding genes. On the other hand, module M10 constitutes in both META and PRO

developmental stages, the second module with more DE coding genes with 46 and 3 genes respectively (Figure 7A).

Meanwhile, *Leishmania donovani* showed high module diversity in differentially expressed ncRNAs, with the modules M4 (24) and M18 (75) modules presenting more representativity of DE ncRNAs (Figure 7B). These modules also presented the major number of DE protein coding genes, with 160 and 87 respectively (Figure 7B).

When considering the presence of differentially expressed coding genes and ncRNAs in the modules identified for *Leishmania major*, we observed that the module M15 presented the largest number of representatives of both ncRNAs (326) and coding genes (408). Additionally, we found the module M2 as the second most represented with DE coding genes with 36 in total. Whereas module M6 was the second most represented with DE ncRNAs, with 38 in total.

In order to better understand the relation between differentially expressed genes along developmental stages and its conservation within *Leishmania* species, we grouped the presence-absence of each coding gene and ncRNA according to its respective expression modules. Surprisingly, we found only 4 differentially ncRNAs shared between *L. braziliensis* and *L. major* linked to the same orthologous coding genes in the co-expression module analyses from both species. Those transcripts (both coding and non-coding) presented similar expression patterns in the same developmental stage in both species. As we showed in Table 3, the gene H2BFM (Histone 2B) was identified as expressed similarly in both *L. major* and *L. braziliensis* in META developmental stage. This gene presented a negatively co-expression pattern associated with two orthologs cis-regulatory ncRNAs in *L. braziliensis* (LbraM2903_32_543492_543513) and *L. major* (LmajFriedlin_33_539769_539790). This result indicate a putative association between the expression of these two cis-regulatory RNAs and the regulation of chromatin structure in these parasites during the META developmental stage. Thus, the expression of those two cis-regulatory RNAs were

385 correlated with the expression of surface -proteophosphoglycan ppg3-coding gene,
 386 suggesting a potential role in pathogenesis for these ncRNAs. Similarly, we found that the
 387 miRNAs LbraM2903_32_716568_716588 (*L. braziliensis*) and
 388 LmajFriedlin_32_1478026_1478046 (*L. major*) co-express negatively with cyclophilin, a
 389 protein that act as chaperone and have a role in protein folding (52,53). These results
 390 denote the importance of different classes of ncRNAs in a variety of physiological processes
 391 in *Leishmania* parasites. Other positive and negative associations between conserved
 392 ncRNAs and coding genes within different species can be observed in Table 3.

Table 3. Expression, conservation, and module relationship between ncRNAs and coding genes

ncRNA	Log ₂ FC	Class	Module	Coding	Log ₂ FC	Name	Weight
LbraM2903_32_543492_543513	2.72	Cis-reg	M3	LbrM.19.0340	-1.59	H2BFM	0.05
				LbrM.22.0710	1.89	zinc finger containing protein	0.04
				LbrM.34.0610	2.74	ppg3	0.04
				LbrM.34.4360	2.11	4E-interacting protein	0.05
LmajFriedlin_33_539769_539790	1.81	Cis-reg	M15	LmjF.09.1340	-5.15	H2BFM	0.41
				LmjF.22.0750	2.14	hypothetical protein	0.37
				LmjF.35.0500	3.50	ppg3	0.43
				LmjF.35.3980	1.51	4E-interacting protein	0.36
LbraM2903_34_1251159_1251179	2.65	Cis-reg	M3	LbrM.19.03401	-1.59	H2BFM	0.04
				LbrM.22.0710	1.89	zinc finger containing protein	0.03
				LbrM.34.0610	2.74	ppg3	0.03
				LbrM.34.4360	2.11	4E-interacting protein	0.04
LmajFriedlin_16_143106_143126	1.66	Cis-reg	M15	LmjF.22.0750	2.14	hypothetical protein	0.26
				LmjF.35.0500	3.50	ppg3	0.43
				LmjF.35.3980	1.51	4E-interacting protein	0.27

				<i>LmjF.09.1340</i>	-5.15	H2BFM	0.35
LbraM2903_723_1367_1388	2.64	Cis-reg	M3	<i>LbrM.19.0340</i>	-1.59	H2BFM	0.04
				<i>LbrM.22.0710</i>	1.89	zinc finger containing protein	0.03
				<i>LbrM.34.0610</i>	2.74	ppg3	0.03
				<i>LbrM.34.4360</i>	2.11	4E-interacting protein	0.04
LmajFriedlin_7_144414_144435	3.41	Cis-reg	M15	<i>LmjF.22.0750</i>	2.14	hypothetical protein	0.26
				<i>LmjF.35.0500</i>	3.50	ppg3	0.43
				<i>LmjF.35.3980</i>	1.51	4E-interacting protein	0.27
				<i>LmjF.09.1340</i>	-5.15	H2BFM	0.35
LbraM2903_32_716568_716588	2.64	miRNA	M3	<i>LbrM.19.0340</i>	-1.59	H2BFM	0.04
				<i>LbrM.22.0710</i>	1.89	hypothetical protein	0.03
				<i>LbrM.34.0610</i>	2.74	ppg3	0.03
				<i>LbrM.34.4360</i>	2.11	4E-interacting protein	0.04
				<i>LbrM.33.1900</i>	-1.61	CYP4A	0.05
LmajFriedlin_32_1478026_1478046	1.71	miRNA	M15	<i>LmjF.22.0750</i>	2.14	hypothetical protein	0.26
				<i>LmjF.35.0500</i>	3.50	ppg3	0.43
				<i>LmjF.35.3980</i>	1.51	4E-interacting protein	0.27
				<i>LmjF.09.1340</i>	-5.15	H2BFM	0.35
				<i>LmjF.25.0910</i>	-1.90	CYP4A	0.35

393

394 DISCUSSION

395 Here we have described for the first time a genome-wide prediction of non-coding RNAs in
396 16 species from the *Leishmania* genus. Our results on 26 different strains revealed
397 expression patterns of coding and non-coding RNAs highly related to different parasite
398 developmental stages. These findings may suggest a role in the regulation of morphological
399 differentiation in both insect and mammal host stages, as occurs in other Trypanosomatidae

(54). So far, different efforts were performed in order to describe the repertoire of distinct ncRNA classes in *Leishmania* parasites (36–38,40,41,43,44). However, different from these works, our computational approach did not focus on any particular ncRNA class, allowing an unbiased genome-wide identification of the ncRNAs repertoire. We followed a similar approach implemented before by our group to describe the sets of ncRNAs available in *Leishmania braziliensis* (43).

We combined sequence similarity searches and covariance model comparisons in order to identify and annotate a large number of ncRNAs in all studied genomes. Additionally, we used publicly available RNA-seq assays to obtain transcriptional evidence that these predictions are generating real transcripts. Our work determined a majority of siRNA-like with length varying between 20 to 25 nt, which coincides to the previously described for *L. braziliensis* (37). Additionally, as reported in *L. major*, most miRNA-like elements possess a length varying around 20 to 26 nt (38). Given our results shown heterogeneity of size and number of predicted ncRNA observed in all species, we identified a median size of 23 nt for all our putative ncRNAs. This contrasts greatly with the data presented by Ruy and colleagues., which in their search using *L. braziliensis* they observed a median size of 281 nt. Interestingly, when comparing our methodology against their methods for ncRNA search, we realized that their methodology was based on the identification of non-coding transcripts and then asked for RNA class, instead we based our search on sequence alignment and probabilistic models. Ruy et al also found more lncRNAs than our approach and our results showed an enrichment in small ncRNAs (44). It suggests that they were identifying primary transcripts that could be precursors of small RNA classes identified in our analysis.

The conservation of ncRNA sequences has been observed in both distant species, such as *Caenorhabditis elegans* and human (55), and more closely related species such as *Arabidopsis thaliana* and *A. lyrata* (56). This conservation has been firmly related to development and diseases (57), or to stress response (58,59). Considering the accepted

426 phylogenetic relationship between the *Leishmania* genus and their potential to infect different
 427 hosts, as well as the ability of this parasite to generate many clinical manifestations, we
 428 studied a possible relationship between sequence conservation at different ncRNA classes
 429 and their role in parasite development by comparing their expression profile in different
 430 developmental stages. Here, we yield evidence of poor conservation of ncRNAs in
 431 *Leishmania* parasites among different spp. Nonetheless, we observed that conservation
 432 increases as phylogenetic distance decreases (Supplementary Figure 2). We observed an
 433 increase in ncRNA conservation in species that are canonically related to the same disease
 434 phenotype (Figure 2), as previously reported for other trypanosomatids (60) and *Leishmania*
 435 parasites (39,44). Sequence conservation is also an indicator of selective pressure, as
 436 occurs with other species, and it can be associated with shared biological processes within
 437 related organisms (61). Previous studies on other pathogens showed that many ncRNA
 438 classes are related to different developmental steps (62,63). Thus, we observed that
 439 conservation of ncRNAs indicates a crucial role in the development and pathogenesis of
 440 *Leishmania* parasites. .

441 In order to understand the differences in gene expression along distinct developmental
 442 stages in *Leishmania* parasites, we selected three representative strains according to their
 443 disease type, subgenus and availability of RNA-seq datasets in public databases. In this
 444 sense we compared the expression of both coding genes and ncRNAs from *L. (Viannia)*
 445 *braziliensis*, which causes cutaneous and mucocutaneous leishmaniasis; *L. (Leishmania)*
 446 *donovani*, which causes visceral leishmaniasis; and *L. (L.) major*, that causes cutaneous
 447 leishmaniasis. In these three strains we found co-expression patterns in both coding and
 448 non-coding RNAs which are exclusively related to specific stages of the developmental cycle
 449 of these parasites. We used this information to associate each cluster of ncRNAs to putative
 450 functions in *Leishmania* spp. different developmental stages (AMA, PRO and META), based
 451 on a co-expression network analysis previously reported (64,65). To date, only one work

employed co-expression networks in order to understand transcriptomic changes through the life cycle of *Leishmania* species, and it was exclusively focused on *L. major* (33) and did not explore ncRNAs. Other approaches suggested possible associations between ncRNAs and coding genes. For instance, Aoki and colleagues obtained functional associations between lncRNAs with Arginine metabolism in *L. amazonensis* (42), whereas a work based on *L. braziliensis* revealed differential expression of several coding and non-coding genes associated with different cycle stages (44). In our study, we provide evidence for the functional annotation of 1,024 ncRNAs divided into these three species of *Leishmania* parasites based on their connectivity with coding genes. Interestingly, we identified that some of these ncRNAs were differentially expressed in specific developmental stages in all three species, but only a few of these were shared between all species.

We identified hundreds of ncRNAs shared within all species (core RNAs). Some of them were characterized as differentially expressed in the same way in co-expressed modules correlated to shared developmental stages. This result led us to evaluate which protein-coding genes were co-expressed to these ncRNAs, as well as to explore which are functions of these coding genes, and if these connections (coding gene-ncRNA) were conserved with orthologous groups in the different species. Based on this, we found that miRNA LbraM2903_32_716568_716588 and their ortholog in *L. major* (miRNA LmajFriedlin_32_1478026_1478046) were both overexpressed in META related modules in both *L. braziliensis* and *L. major*, respectively. These two ncRNAs were linked with five coding genes conserved in both *L. braziliensis* and *L. major*, of which 3 were overrepresented in META and 2 PRO developmental stages in both species.

Also, we identified that four of these five co-expressed coding genes were co-expressed with other two cis-regulatory ncRNAs in both *L. braziliensis* and *L. major*. This could suggest that these genes might be actively regulated by the action of these ncRNAs. These results demonstrated that those ncRNAs may play a crucial role in *Leishmania* parasite

development. The relationship of these ncRNAs with different coding genes demonstrated a similar expression behavior shared within *Leishmania* species. For instance, in the comparison between META and PRO for *L. braziliensis* and *L. major*, we noticed the presence of a gene that encodes for histone H2B. Several studies have shown that histone proteins are evolutionarily diversified in trypanosomatids in comparison with histones from other eukaryotes (26,66,67). On other hand, several works demonstrated that histone genes, specially histones H2B and H4 show antigenicity and were recognized by antibodies (68) and when treated with Sodium antimony gluconate (SAG) showed an overexpression of these proteins (69). In a recent work, demonstrated that histone H3 could be exported to the host nucleus and bind with chromatin, which would indicate that *Leishmania* histones not only play a role in the regulation of parasite chromatin, but are also important for the survival of the parasite in the host cells, through changes in the structure and dynamics of the host chromatin (66). The presence of anti-H2B antibodies has been found in symptomatic visceral and cutaneous leishmaniasis patients (70,71) corroborating that histones are needed for the correct development and survival of the parasite. Additionally, this indicates that ncRNAs that coexpress with histone proteins may play an important role in the regulation of development of the parasite and survival in hosts by the regulation of histones and other proteins. Thus, we propose histone H2B and the co-expressed ncRNAs as possible markers of drug susceptibility and survival.

4E-interacting protein (4E-IP), is known to be related to translation protein (72). Recently in a work developed by Mabile and cols demonstrated that 4E-IP is essential for development and infectivity to vertebrate hosts in *Trypanosoma brucei* (73). Besides this protein is constitutively expressed in *Leishmania* (74), another report demonstrates that 4E-IP is essential for promastigote stage-specific translation (72). This indicates the need for regulation in amastigote developmental stage. Our evidence shows the positive coexpression of 4E-IP and several ncRNAs in metacyclic promastigote stage; however,

504 further analysis is required to identify how the expression of these ncRNAs could affect the
505 expression of 4E-IP in amastigote developmental stage.

506 Surface protein-encoding gene proteophosphoglycan ppg3 has been reported as virulence
507 factor in *Leishmania* parasites, nonetheless the downregulation of this protein in *Leishmania*
508 *amazonensis* has been related with resistant phenotype in presence of trivalent sodium
509 stibogluconate (75). Our results reveals the positive correlation of ppg3 with different
510 conserved ncRNAs expression, which leads us to propose these ncRNAs as possible
511 regulators of ppg3 expression in *Leishmania* parasites

512 Cyclophilin A (CYPA) is a protein that acts as chaperone and has a role in protein folding
513 (52,53), but also is a regulator of immune response and inflammatory reaction in vertebrates
514 (76). On the other hand, CYPA has a role in the life cycle and pathogenesis of several
515 parasites, such as *Leishmania*, affecting the replication, infection and survival in host cells
516 (77). In *Trypanosoma brucei* cyclophilin A was shown to be essential for parasite
517 proliferation and susceptibility of the host (78). In a study from do Monte-Neto and
518 colleagues, found that CYPA is overexpressed in *Leishmania amazonensis* resistant to
519 antimony treatment (79). Demonstrating the role of this protein in the pathogenesis of the
520 parasite, as well as changes in its regulation can trigger resistance to this conventional
521 treatment. Our results showed that the expression of CYPA in wild type conditions is
522 downregulated in META developmental stage in both, *L. braziliensis* and *L. major*, meanwhile
523 it presents a negative coexpression with a conserved miRNA in both species, which
524 indicates the possible role of this miRNA in the post-transcriptional regulation of CYPA;
525 however more analysis must be performed to corroborate this effect of this miRNA in the
526 regulation of CYPA.

527 In summary, our work is the first to identify novel ncRNAs in 26 genome isolates from 16
528 *Leishmania* species. Additionally, it revealed a new set of non-coding genes that are
529 involved in different developmental stages in the parasite life cycle and may play an

important role in infection and survival. Furthermore, the co-expression profile of coding and non-coding genes functional insights for the potential role of hundreds of ncRNAs through guilt-by-association inferences. Finally, our results provide evidence of possible mechanisms underlying co-regulation between coding-ncRNA interaction, mainly the role of ncRNAs in the regulation histones interaction. This opened the way for further research on the role of ncRNAs and their possible relationship with parasite survival.

METHODS

Databases and datasets

All genome sequences of 26 strains from 16 different species of *Leishmania* spp. were downloaded from NCBI (80) FTP site and TriTrypDB release 35 (81) in October 2019 (Supplementary Table 1). RNA sequencing (RNA-seq) raw data from three studies originated through Illumina platform were downloaded from SRA NCBI (82) in October 2019 (Supplementary Table 2), resulting in 30 biological samples from three *Leishmania* species *L. braziliensis* M2903 (MHOM/BR/75/M2903) and *L. braziliensis* M2904 (MHOM/BR/75/M2904), *L. donovani* BPK282A1 (MHOM/NP/02BPK282) and *L. major* Friedlin (MHOM/IL/81/Friedlin). GFF format annotation files, when available, were also downloaded from TriTrypDB release 35 on the same date.

Predicting the repertoire of ncRNAs in *Leishmania* spp genomes

The predictions of ncRNAs in *Leishmania* spp. genomes were carried out using two different schemes. The first approach is based on sequence homology detection, using the set of non-redundant sequences available at the NR2 database (46). For this, we employed an in-house developed script to perform the search. This script takes as input a FASTA file with all sequences from NR2 and a genome index made with Bowtie2 (83) for each genome, generating as output a BED file for each chromosome with all mapped ncRNAs. The other

ncRNA prediction scheme was based on covariance models search, as implemented in StructRNAFinder (84), using Rfam 14.1 (47) known RNA families. An e-value cut-off of 0.001 was applied for cmsearch and a score of 10 was used to identify and annotate RNA families with StructRNAfinder, as previously applied for *L. braziliensis* (43). The two BED files obtained from both sequence homology and covariance models approaches were merged, in order to eliminate redundancies, using the BEDtools merge (85), with an overlapping cut-off of 10 nt. The final set of ncRNA FASTA sequences was obtained using the BEDtools getfasta (85). The final RNA class annotation for each predicted ncRNA was obtained by using the consensus made between the classes that were mapped by both searches.

565

566 **Transcriptional evidence of predicted ncRNAs**

Once the putative ncRNAs were identified, their expression was used to validate their existence. For this, all RNA-seq files deposited in SRA (82) corresponding to *Leishmania* species were downloaded in October 2019. These files were then separated by specific strains and only those experiments corresponding to the strains with ncRNA predictions were preserved. We selected *Leishmania braziliensis* MHOM/BR/1975/M2903, *Leishmania donovani* BPK282A1 (MHOM/NP/2020/BPK282A1) and *Leishmania major* Friedlin (MHOM/IL/1981/Friedlin) for subsequent analyzes, according to previously described parameters (39). Next, we employed FastQC v0.11.8 (86) and Trimmomatic 0.36 (87) to evaluate and filter out low quality reads, considering a Phred cut-off value of Q = 30. Then, we mapped the remaining high quality reads using Bowtie2 (85) to each one of the three selected genomes. Expression read count values for all predicted ncRNAs were calculated using HTSeq-count 0.7.2 (88) and normalized to CPM using CORAZON (89). A CPM cut-off value of 0.5 was employed to determine if a particular ncRNA prediction was deemed to exist in each genome.

581

582 **Comparative analysis of ncRNAs in *Leishmania* spp.**

583 To compare the ncRNA repertoire in *Leishmania* spp., we analyzed the conservation of all
584 predicted ncRNA on all studied genomes. To do this, we recovered the FASTA sequence for
585 all ncRNA obtained in the prediction step and grouped these sequences using CD-HIT
586 version 4.8.1 (90), considering 90% of sequence similarity and coverage of both largest and
587 shortest sequences to classify a ncRNA as a cluster member. A binary presence-absence
588 matrix was then constructed based on each cluster of predicted ncRNAs. The set of ncRNA
589 clusters shared within all strains was defined as the “core-RNAome”, meanwhile the sum of
590 the “accessory-RNAome” (strain specific and auxiliary ncRNA clusters) and the
591 core-RNAome, was defined as the “pan-RNAome”.

592 In order to visualize the species clustering according to the ncRNA family distribution, we
593 applied a hierarchical clustering with Euclidean distance using the presence/absence matrix.
594 Next, we compared the generated cladogram with a phylogenetic tree obtained from a set of
595 conserved orthologous proteins. For that, core-genome phylogenetic trees were built using
596 613 conserved orthologous proteins from the core-genome of 26 *Leishmania*
597 representatives. For this, we employed a bidirectional best hit algorithm for orthologs
598 clustering using ProteinOrtho v5.11 (91) with default parameters. Next, conserved
599 orthologous multiprotein family sequences were aligned using MAFFT version v7.453 (92),
600 considering the L-INS-I iterative refinement method. The alignments were masked to remove
601 unreliable aligned regions with GBLOCKS version 0.91b (93). Maximum likelihood trees
602 were prepared for concatenated alignments through IQ-TREE version 1.6.12 (94), using
603 1000 replicates as bootstrap with best-suited substitution model. The final tree was
604 visualized using Figtree (<http://tree.bio.ed.ac.uk/software/figtree/>).

605

606 **Co-expression network analysis of developmental stages based on coding genes and** 607 **non-coding RNAs expression patterns**

We employed a guilty-by-association approach to generate co-expression modules and networks associating ncRNAs and coding genes to each developmental stage, through WGCNA tool version 1.69 (50). An appropriate soft-threshold power was selected based on the pickSoft threshold function of WGCNA, and each soft threshold was used to generate a scale-free topology plot. The cluster dendrogram plot and clustering tree of co-expression gene modules were created with cut-offs for cutHeight = 0.25 and minSize = 30. After that, the module eigengene (ME) was employed to indicate the expression profiles of each module.

To analyze the correlation of each module and different *Leishmania* spp. developmental stages we used a “Module-development stage relationship” association, taking as input the expression data of different developmental stages and the modules correlations. For that, we used the module eigengenes function of WGCNA, and all module with p -value < 0.05 and $r > |0.75|$ were considered as statistically highly correlated. Next, to identify genes presenting high correlation to different developmental stages, we applied gene significance (GS) and module membership (MM) measures, by assigning a gene X as highly related to developmental stage Z , if an X^Z GS ≥ 0.70 , $p \leq 0.05$ and a MM $\geq |0.5|$, $p \leq 0.05$.

624

625 **Functional associations of ncRNA-related modules in *Leishmania* parasites**

To obtain functional associations for each ncRNA we employed a module based method based on the co-expression networks generated in the previous topic. We applied a modified version of the method described by Liao and cols (64). First, we obtained the significant co-expressed modules of the previous step and then filtered those modules by weight, considering only those links between coding and non-coding genes in which their weight were between ≥ 0.09 for *L. donovani* and *L. major* and ≥ 0.35 for *L. braziliensis*. Gene Ontology (GO) enrichment was calculated using Goseq (51) with FDR adjusted p -values to

633 select only those terms with a p -value below 0.05. To facilitate the analysis interpretation GO
634 terms were fused into their level-3 ancestors in the hierarchy.

635

636 **Inferring orthologous groups per modules across developmental stages in**
637 ***Leishmania* spp.**

638 Each highly significant module associated with a particular developmental stage from one
639 species had their protein coding genes cross-referenced with their orthologous genes
640 available in the module enriched in the same stage in the other species. For that, we
641 employed OrthoVenn2 (95) with an e-value of $1e-5$ as cutoff for all-to-all protein similarity
642 comparisons. Once orthologous clusters were obtained, we performed GO enrichment
643 analysis for all shared clusters. For the ncRNAs, the same method to construct the
644 pan-RNAome was used to determine their modules representation across species.

645

646 **Differential expression analysis among developmental stages in *Leishmania* spp.**

647 The expression values obtained when determining the transcriptional evidence of coding
648 genes and ncRNAs were used for differential expression analysis. DESeq2 (96) was used in
649 order to determine statistically differentially expressed (DE) coding and non-coding RNAs. To
650 consider a gene as DE we used an empirical probability distribution $|\log_2 \text{fold-change}|$ with a
651 threshold > 1.5 and a p value < 0.05 .

652

653 **ACKNOWLEDGMENTS**

654 We acknowledge the kindly support received from Dr. Evandro Ferrada, Dr Fernando Alfaro
655 and Dr Roberto Mercado. Their patience and useful discussions were valuable for the
656 improvement of our work. Also, we would like to thank the members of both NBL and LIB
657 laboratories for their constructive criticism and advice along the development of this
658 manuscript.

659

660 REFERENCES

- 661 1. Murray HW, Berman JD, Davies CR, Saravia NG. Advances in leishmaniasis [Internet].
662 Vol. 366, The Lancet. 2005. p. 1561–77. Available from:
663 [http://dx.doi.org/10.1016/s0140-6736\(05\)67629-5](http://dx.doi.org/10.1016/s0140-6736(05)67629-5)
- 664 2. Burza S, Croft SL, Boelaert M. Leishmaniasis. Lancet. 2018 Sep
665 15;392(10151):951–70.
- 666 3. Reithinger R, Dujardin J-C, Louzir H, Pirmez C, Alexander B, Brooker S. Cutaneous
667 leishmaniasis. Lancet Infect Dis. 2007 Sep;7(9):581–96.
- 668 4. Subramanian A, Sarkar RR. Perspectives on Leishmania Species and Stage-specific
669 Adaptive Mechanisms. Trends Parasitol. 2018 Dec;34(12):1068–81.
- 670 5. WHO | Leishmaniasis. 2020 Jan 13 [cited 2020 Feb 26]; Available from:
671 <http://www.who.int/leishmaniasis/en/>
- 672 6. Bates PA. Revising Leishmania's life cycle [Internet]. Vol. 3, Nature Microbiology. 2018.
673 p. 529–30. Available from: <http://dx.doi.org/10.1038/s41564-018-0154-2>
- 674 7. Podinovskaia M, Descoteaux A. Leishmania and the macrophage: a multifaceted
675 interaction. Future Microbiol. 2015;10(1):111–29.
- 676 8. Nieto A, Domínguez-Bernal G, Orden JA, De La Fuente R, Madrid-Elena N, Carrión J.
677 Mechanisms of resistance and susceptibility to experimental visceral leishmaniasis:
678 BALB/c mouse versus Syrian hamster model. Vet Res. 2011 Feb 23;42:39.
- 679 9. Rodríguez AE, Estévez JO, Nevot MC, Barrios A, Florin-Christensen M. Leishmania
680 [Internet]. Parasitic Protozoa of Farm Animals and Pets. 2018. p. 289–311. Available
681 from: http://dx.doi.org/10.1007/978-3-319-70132-5_13
- 682 10. Dostálová A, Volf P. Leishmania development in sand flies: parasite-vector interactions
683 overview. Parasit Vectors. 2012 Dec 3;5:276.
- 684 11. Maran SR, de Lemos Padilha Pitta JL, Dos Santos Vasconcelos CR, McDermott SM,
685 Rezende AM, Silvio Moretti N. Epitranscriptome machinery in Trypanosomatids: New
686 players on the table? Mol Microbiol [Internet]. 2021 Jan 29; Available from:
687 <http://dx.doi.org/10.1111/mmi.14688>
- 688 12. Haile S, Papadopoulou B. Developmental regulation of gene expression in
689 trypanosomatid parasitic protozoa. Curr Opin Microbiol. 2007 Dec;10(6):569–77.
- 690 13. Shimoni Y, Friedlander G, Hetzroni G, Niv G, Altuvia S, Biham O, et al. Regulation of
691 gene expression by small non-coding RNAs: a quantitative view [Internet]. Vol. 3,
692 Molecular Systems Biology. 2007. p. 138. Available from:
693 <http://dx.doi.org/10.1038/msb4100181>
- 694 14. Balasubramanian D, Vanderpool CK. New developments in post-transcriptional
695 regulation of operons by small RNAs. RNA Biol. 2013 Mar;10(3):337–41.
- 696 15. Vaucheret H. Post-transcriptional small RNA pathways in plants: mechanisms and

- 697 regulations. *Genes Dev.* 2006 Apr 1;20(7):759–71.
- 698 16. Sun BK, Tsao H. Small RNAs in development and disease [Internet]. Vol. 59, *Journal of*
699 the American Academy of Dermatology. 2008. p. 725–37. Available from:
700 <http://dx.doi.org/10.1016/j.jaad.2008.08.017>
- 701 17. Romano G, Veneziano D, Acunzo M, Croce CM. Small non-coding RNA and cancer.
702 *Carcinogenesis.* 2017 May 1;38(5):485–91.
- 703 18. Shirahama S, Miki A, Kaburaki T, Akimitsu N. Long Non-coding RNAs Involved in
704 Pathogenic Infection. *Front Genet.* 2020 May 26;11:454.
- 705 19. Zhang P, Wu W, Chen Q, Chen M. Non-Coding RNAs and their Integrated Networks
706 [Internet]. Vol. 16, *Journal of Integrative Bioinformatics.* 2019. Available from:
707 <http://dx.doi.org/10.1515/jib-2019-0027>
- 708 20. Fu X-D. Non-coding RNA: a new frontier in regulatory biology. *Natl Sci Rev.* 2014 Jun
709 1;1(2):190–204.
- 710 21. Koerner MV, Pauler FM, Huang R, Barlow DP. The function of non-coding RNAs in
711 genomic imprinting [Internet]. Vol. 136, *Development.* 2009. p. 1771–83. Available from:
712 <http://dx.doi.org/10.1242/dev.030403>
- 713 22. Romero-Barrios N, Legascue MF, Benhamed M, Ariel F, Crespi M. Splicing regulation by
714 long noncoding RNAs. *Nucleic Acids Res.* 2018 Mar 16;46(5):2169–84.
- 715 23. Amaral PP, Leonardi T, Han N, Viré E, Gascoigne DK, Arias-Carrasco R, et al. Genomic
716 positional conservation identifies topological anchor point RNAs linked to developmental
717 loci. *Genome Biol.* 2018 Mar 15;19(1):32.
- 718 24. Holley CL, Topkara VK. An introduction to small non-coding RNAs: miRNA and
719 snoRNA. *Cardiovasc Drugs Ther.* 2011 Apr;25(2):151–9.
- 720 25. Chacko N, Lin X. Non-coding RNAs in the development and pathogenesis of eukaryotic
721 microbes. *Appl Microbiol Biotechnol.* 2013 Sep;97(18):7989–97.
- 722 26. Ivens AC, Peacock CS, Worthey EA, Murphy L, Aggarwal G, Berriman M, et al. The
723 genome of the kinetoplastid parasite, *Leishmania major*. *Science.* 2005 Jul
724 15;309(5733):436–42.
- 725 27. Peacock CS, Seeger K, Harris D, Murphy L, Ruiz JC, Quail MA, et al. Comparative
726 genomic analysis of three *Leishmania* species that cause diverse human disease. *Nat*
727 *Genet.* 2007 Jul;39(7):839–47.
- 728 28. Rogers MB, Hilley JD, Dickens NJ, Wilkes J, Bates PA, Depledge DP, et al.
729 Chromosome and gene copy number variation allow major structural change between
730 species and strains of *Leishmania*. *Genome Res.* 2011 Dec;21(12):2129–42.
- 731 29. Downing T, Imamura H, Decuypere S, Clark TG, Coombs GH, Cotton JA, et al. Whole
732 genome sequencing of multiple *Leishmania donovani* clinical isolates provides insights
733 into population structure and mechanisms of drug resistance. *Genome Res.* 2011
734 Dec;21(12):2143–56.
- 735 30. Real F, Vidal RO, Carazzolle MF, Mondego JMC, Costa GGL, Herai RH, et al. The
736 genome sequence of *Leishmania (Leishmania) amazonensis*: functional annotation and

- 737 extended analysis of gene models. *DNA Res.* 2013 Dec;20(6):567–81.
- 738 31. Zhang W-W, Matlashewski G. Screening *Leishmania donovani*-specific genes required
739 for visceral infection. *Mol Microbiol.* 2010 Jul;77(2):505–17.
- 740 32. Camacho E, Rastrojo A, Sanchiz Á, la Fuente SG, Aguado B, Requena JM. *Leishmania*
741 *Mitochondrial Genomes: Maxicircle Structure and Heterogeneity of Minicircles* [Internet].
742 Vol. 10, *Genes*. 2019. p. 758. Available from: <http://dx.doi.org/10.3390/genes10100758>
- 743 33. Inbar E, Keith Hughitt V, Dillon LAL, Ghosh K, El-Sayed NM, Sacks DL. The
744 Transcriptome of *Leishmania major* Developmental Stages in Their Natural Sand Fly
745 Vector [Internet]. Vol. 8, *mBio*. 2017. Available from:
746 <http://dx.doi.org/10.1128/mbio.00029-17>
- 747 34. Butenko A, Kostygov AY, Sádlová J, Kleschenko Y, Bečvář T, Podešvová L, et al.
748 Comparative genomics of *Leishmania* (*Mundinia*). *BMC Genomics*. 2019 Oct
749 11;20(1):726.
- 750 35. Kumar A, Pandey SC, Samant M. DNA-based microarray studies in visceral
751 leishmaniasis: identification of biomarkers for diagnostic, prognostic and drug target for
752 treatment. *Acta Trop.* 2020 May 8;208:105512.
- 753 36. Dumas C, Chow C, Müller M, Papadopoulou B. A novel class of developmentally
754 regulated noncoding RNAs in *Leishmania*. *Eukaryot Cell.* 2006 Dec;5(12):2033–46.
- 755 37. Atayde VD, Shi H, Franklin JB, Carriero N, Notton T, Lye L-F, et al. The structure and
756 repertoire of small interfering RNAs in *Leishmania* (*Viannia*) *braziliensis* reveal
757 diversification in the trypanosomatid RNAi pathway. *Mol Microbiol.* 2013
758 Feb;87(3):580–93.
- 759 38. Sahoo G, Ansari M, Dikhit M, Gupta N, Rana S, Das P. Computational Identification of
760 microRNA-like Elements in *Leishmania major* [Internet]. Vol. 2, *MicroRNA*. 2014. p.
761 225–30. Available from: <http://dx.doi.org/10.2174/2211536602666131203232422>
- 762 39. Lambert U, Oviedo Ovando ME, Vasconcelos EJ, Unrau PJ, Myler PJ, Reiner NE.
763 Small RNAs derived from tRNAs and rRNAs are highly enriched in exosomes from both
764 old and new world *Leishmania* providing evidence for conserved exosomal RNA
765 Packaging. *BMC Genomics*. 2015 Mar 5;16:151.
- 766 40. Eliaz D, Doniger T, Tkacz ID, Biswas VK, Gupta SK, Kolev NG, et al. Genome-wide
767 analysis of small nucleolar RNAs of *Leishmania major* reveals a rich repertoire of RNAs
768 involved in modification and processing of rRNA. *RNA Biol.* 2015 May
769 13;12(11):1222–55.
- 770 41. Castro FF, Ruy PC, Zeviani KN, Santos RF, Toledo JS, Cruz AK. Evidence of putative
771 non-coding RNAs from *Leishmania* untranslated regions [Internet]. Vol. 214, *Molecular*
772 *and Biochemical Parasitology*. 2017. p. 69–74. Available from:
773 <http://dx.doi.org/10.1016/j.molbiopara.2017.04.002>
- 774 42. Aoki JI, Muxel SM, Zampieri RA, Laranjeira-Silva MF, Müller KE, Nerland AH, et al.
775 RNA-seq transcriptional profiling of *Leishmania amazonensis* reveals an
776 arginase-dependent gene expression regulation. *PLoS Negl Trop Dis.* 2017
777 Oct;11(10):e0006026.

- 778 43. Torres F, Arias-Carrasco R, Caris-Maldonado JC, Barral A, Maracaja-Coutinho V, De
779 Queiroz ATL. LeishDB: a database of coding gene annotation and non-coding RNAs in
780 *Leishmania braziliensis*. Database [Internet]. 2017 Jan 1;2017. Available from:
781 <http://dx.doi.org/10.1093/database/bax047>
- 782 44. Ruy PDC, De Cássia Ruy P, Monteiro-Teles NM, Magalhães RDM, Freitas-Castro F,
783 Dias L, et al. Comparative transcriptomics in *Leishmania braziliensis*: disclosing
784 differential gene expression of coding and putative noncoding RNAs across
785 developmental stages [Internet]. Vol. 16, RNA Biology. 2019. p. 639–60. Available from:
786 <http://dx.doi.org/10.1080/15476286.2019.1574161>
- 787 45. Guegan F, Bento F, Neves D, Sequeira M, Notredame C, Figueiredo LM. A long
788 non-coding RNA controls parasite differentiation in African trypanosomes [Internet].
789 Available from: <http://dx.doi.org/10.1101/2020.05.03.074625>
- 790 46. Paschoal AR, Maracaja-Coutinho V, Setubal JC, Simões ZLP, Verjovski-Almeida S,
791 Durham AM. Non-coding transcription characterization and annotation [Internet]. Vol. 9,
792 RNA Biology. 2012. p. 274–82. Available from: <http://dx.doi.org/10.4161/rna.19352>
- 793 47. Kalvari I, Nawrocki EP, Argasinska J, Quinones-Olvera N, Finn RD, Bateman A, et al.
794 Non-Coding RNA Analysis Using the Rfam Database. Curr Protoc Bioinformatics. 2018
795 Jun;62(1):e51.
- 796 48. Bi K, Chen Y, Zhao S, Kuang Y, John Wu C-H. Current Visceral Leishmaniasis
797 Research: A Research Review to Inspire Future Study. Biomed Res Int. 2018 Jul
798 10;2018:9872095.
- 799 49. Al-Salem W, Herricks JR, Hotez PJ. A review of visceral leishmaniasis during the
800 conflict in South Sudan and the consequences for East African countries. Parasit
801 Vectors. 2016 Aug 22;9:460.
- 802 50. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network
803 analysis [Internet]. Vol. 9, BMC Bioinformatics. 2008. Available from:
804 <http://dx.doi.org/10.1186/1471-2105-9-559>
- 805 51. Young MD, Wakefield MJ, Smyth GK, Oshlack A. Gene ontology analysis for RNA-seq:
806 accounting for selection bias. Genome Biol. 2010 Feb 4;11(2):R14.
- 807 52. Kofron JL, Kuzmic P, Kishore V, Colón-Bonilla E, Rich DH. Determination of kinetic
808 constants for peptidyl prolyl cis-trans isomerases by an improved spectrophotometric
809 assay. Biochemistry. 1991 Jun 25;30(25):6127–34.
- 810 53. Barik S. Immunophilins: for the love of proteins. Cell Mol Life Sci. 2006
811 Dec;63(24):2889–900.
- 812 54. Michaeli S. Non-coding RNA and the complex regulation of the trypanosome life cycle.
813 Curr Opin Microbiol. 2014 Aug;20:146–52.
- 814 55. Ibáñez-Ventoso C, Vora M, Driscoll M. Sequence relationships among *C. elegans*, *D.*
815 *melanogaster* and human microRNAs highlight the extensive conservation of
816 microRNAs in biology. PLoS One. 2008 Jul 30;3(7):e2818.
- 817 56. Jones-Rhoades MW. Conservation and divergence in plant microRNAs. Plant Mol Biol.

- 818 2012 Sep;80(1):3–16.
- 819 57. Tsagakis I, Douka K, Birds I, Aspden JL. Long non-coding RNAs in development and
820 disease: Conservation to mechanisms. *J Pathol.* 2020 Mar 16;8:310433.
- 821 58. Sanz-Carbonell A, Marques MC, Martinez G, Gomez G. Dynamic architecture and
822 regulatory implications of the miRNA network underlying the response to stress in
823 melon. *RNA Biol.* 2020 Feb;17(2):292–308.
- 824 59. Barakat A, Wall PK, Diloroto S, Depamphilis CW, Carlson JE. Conservation and
825 divergence of microRNAs in *Populus*. *BMC Genomics.* 2007 Dec 31;8:481.
- 826 60. Doniger T, Katz R, Wachtel C, Michaeli S, Unger R. A comparative genome-wide study
827 of ncRNAs in trypanosomatids. *BMC Genomics.* 2010 Nov 4;11:615.
- 828 61. Walter Costa MB, Höner Zu Siederdisen C, Dunjić M, Stadler PF, Nowick K. SSS-test:
829 a novel test for detecting positive selection on RNA secondary structure. *BMC*
830 *Bioinformatics.* 2019 Mar 21;20(1):151.
- 831 62. Cai P, Piao X, Hao L, Liu S, Hou N, Wang H, et al. A deep analysis of the small
832 non-coding RNA population in *Schistosoma japonicum* eggs. *PLoS One.* 2013 May
833 14;8(5):e64003.
- 834 63. Mourier T, Carret C, Kyes S, Christodoulou Z, Gardner PP, Jeffares DC, et al.
835 Genome-wide discovery and verification of novel structured RNAs in *Plasmodium*
836 *falciparum*. *Genome Res.* 2008 Feb;18(2):281–92.
- 837 64. Liao Q, Shen J, Liu J, Sun X, Zhao G, Chang Y, et al. Genome-wide identification and
838 functional annotation of *Plasmodium falciparum* long noncoding RNAs from RNA-seq
839 data. *Parasitol Res.* 2014 Apr;113(4):1269–81.
- 840 65. Liao Q, Liu C, Yuan X, Kang S, Miao R, Xiao H, et al. Large-scale prediction of long
841 non-coding RNA functions in a coding–non-coding gene co-expression network. *Nucleic*
842 *Acids Res.* 2011 May 1;39(9):3864–78.
- 843 66. Dacher M, Tachiwana H, Horikoshi N, Kujirai T, Taguchi H, Kimura H, et al. Incorporation
844 and influence of *Leishmania* histone H3 in chromatin. *Nucleic Acids Res.* 2019 Dec
845 16;47(22):11637–48.
- 846 67. Galanti N, Galindo M, Sabaj V, Espinoza I, Toro GC. Histone genes in trypanosomatids.
847 *Parasitol Today.* 1998 Feb;14(2):64–70.
- 848 68. Soto M, Requena JM, Quijada L, Perez MJ, Nieto CG, Guzman F, et al. Antigenicity of
849 the *Leishmania infantum* histones H2B and H4 during canine visceral leishmaniasis
850 leishmaniasis. *Clin Exp Immunol.* 1999 Feb;115(2):342–9.
- 851 69. Singh R, Kumar D, Duncan RC, Nakhasi HL, Salotra P. Overexpression of histone H2A
852 modulates drug susceptibility in *Leishmania* parasites. *Int J Antimicrob Agents.* 2010
853 Jul;36(1):50–7.
- 854 70. Meddeb-Garnaoui A, Toumi A, Ghelis H, Mahjoub M, Louzir H, Chenik M. Cellular and
855 humoral responses induced by *Leishmania* histone H2B and its divergent and
856 conserved parts in cutaneous and visceral leishmaniasis patients, respectively. *Vaccine.*
857 2010 Feb 17;28(7):1881–6.

- 858 71. Carmelo E, Zurita AI, Martinez E, Valladares B. The sera from individuals suffering from
859 cutaneous leishmaniasis due to *Leishmania braziliensis* present antibodies against
860 parasitic conserved proteins, but not their human counterparts. *Parasite*. 2006
861 Sep;13(3):231–6.
- 862 72. Zinoviev A, Léger M, Wagner G, Shapira M. A novel 4E-interacting protein in
863 *Leishmania* is involved in stage-specific translation pathways. *Nucleic Acids Res*. 2011
864 Oct;39(19):8404–15.
- 865 73. Mabile D, Cardoso Santos C, Hendrickx R, Claes M, Takac P, Clayton C, et al. 4E
866 Interacting Protein as a Potential Novel Drug Target for Nucleoside Analogues in
867 *Trypanosoma Brucei*. *Microorganisms*. 2021 Apr 13;9(4):826.
- 868 74. Leifso K, Cohen-Freue G, Dogra N, Murray A, McMaster WR. Genomic and proteomic
869 expression analysis of *Leishmania* promastigote and amastigote life stages: the
870 *Leishmania* genome is constitutively expressed. *Mol Biochem Parasitol*. 2007
871 Mar;152(1):35–46.
- 872 75. Patino LH, Muskus C, Ramírez JD. Transcriptional responses of *Leishmania*
873 (*Leishmania*) *amazonensis* in the presence of trivalent sodium stibogluconate. *Parasit*
874 *Vectors*. 2019 Jul 12;12(1):348.
- 875 76. Qian Z, Yue W, Li B. [Progress of CypA and lung cancer-related research]. *Zhongguo*
876 *Fei Ai Za Zhi*. 2010 Aug;13(8):827–31.
- 877 77. Liao Y, Luo D, Peng K, Zeng Y. Cyclophilin A: a key player for etiological agent infection.
878 *Appl Microbiol Biotechnol*. 2021 Feb;105(4):1365–77.
- 879 78. Pellé R, McOdimba F, Chuma F, Wasawo D, Pearson TW, Murphy NB. The African
880 trypanosome cyclophilin A homologue contains unusual conserved central and
881 N-terminal domains and is developmentally regulated. *Gene*. 2002 May
882 15;290(1-2):181–91.
- 883 79. do Monte-Neto RL, Coelho AC, Raymond F, Légaré D, Corbeil J, Melo MN, et al. Gene
884 expression profiling and molecular characterization of antimony resistance in
885 *Leishmania amazonensis*. *PLoS Negl Trop Dis*. 2011 May;5(5):e1167.
- 886 80. Sayers EW, Beck J, Brister JR, Bolton EE, Canese K, Comeau DC, et al. Database
887 resources of the National Center for Biotechnology Information. *Nucleic Acids Res*.
888 2020 Jan 8;48(D1):D9–16.
- 889 81. Aslett M, Aurrecochea C, Berriman M, Brestelli J, Brunk BP, Carrington M, et al.
890 TriTrypDB: a functional genomic resource for the Trypanosomatidae. *Nucleic Acids Res*.
891 2010 Jan;38(Database issue):D457–62.
- 892 82. Leinonen R, Sugawara H, Shumway M, on behalf of the International Nucleotide
893 Sequence Database Collaboration. The Sequence Read Archive [Internet]. Vol. 39,
894 *Nucleic Acids Research*. 2011. p. D19–21. Available from:
895 <http://dx.doi.org/10.1093/nar/gkq1019>
- 896 83. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods*.
897 2012 Mar 4;9(4):357–9.
- 898 84. Arias-Carrasco R, Vásquez-Morán Y, Nakaya HI, Maracaja-Coutinho V.

- 899 StructRNAfinder: an automated pipeline and web server for RNA families prediction.
900 BMC Bioinformatics. 2018 Feb 17;19(1):55.
- 901 85. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic
902 features [Internet]. Vol. 26, Bioinformatics. 2010. p. 841–2. Available from:
903 <http://dx.doi.org/10.1093/bioinformatics/btq033>
- 904 86. Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput
905 Sequence Data [Internet]. [cited 2020 Feb 26]. Available from:
906 <https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>
- 907 87. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence
908 data. Bioinformatics. 2014 Aug 1;30(15):2114–20.
- 909 88. Anders S, Pyl PT, Huber W. HTSeq--a Python framework to work with high-throughput
910 sequencing data. Bioinformatics. 2015 Jan 15;31(2):166–9.
- 911 89. Ramos TAR, Maracaja-Coutinho V, Ortega JM, do Rêgo TG. CORAZON: a web server
912 for data normalization and unsupervised clustering based on expression profiles. BMC
913 Res Notes. 2020 Jul 14;13(1):338.
- 914 90. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation
915 sequencing data. Bioinformatics. 2012 Dec 1;28(23):3150–2.
- 916 91. Lechner M, Findeiss S, Steiner L, Marz M, Stadler PF, Prohaska SJ. Proteinortho:
917 detection of (co-)orthologs in large-scale analysis. BMC Bioinformatics. 2011 Apr
918 28;12:124.
- 919 92. Katoh K, Misawa K, Kuma K-I, Miyata T. MAFFT: a novel method for rapid multiple
920 sequence alignment based on fast Fourier transform. Nucleic Acids Res. 2002 Jul
921 15;30(14):3059–66.
- 922 93. Castresana J. Selection of Conserved Blocks from Multiple Alignments for Their Use in
923 Phylogenetic Analysis [Internet]. Vol. 17, Molecular Biology and Evolution. 2000. p.
924 540–52. Available from: <http://dx.doi.org/10.1093/oxfordjournals.molbev.a026334>
- 925 94. Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: a fast and effective
926 stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol. 2015
927 Jan;32(1):268–74.
- 928 95. Xu L, Dong Z, Fang L, Luo Y, Wei Z, Guo H, et al. OrthoVenn2: a web server for
929 whole-genome comparison and annotation of orthologous clusters across multiple
930 species. Nucleic Acids Res. 2019 Jul 2;47(W1):W52–8.
- 931 96. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for
932 RNA-seq data with DESeq2 [Internet]. Vol. 15, Genome Biology. 2014. Available from:
933 <http://dx.doi.org/10.1186/s13059-014-0550-8>

934

935 FIGURE CAPTIONS

936 **Figure 1. ncRNAs repertoire and transcriptional evidence in *Leishmania* parasites. A.**

937 The frequency of all predicted ncRNAs for each strain of *Leishmania* spp. evaluated. Those
 938 ncRNAs representing less than 0.5% were categorized as others. We can emphasize the
 939 presence mainly of Cis-regulatory RNA, these RNAs are elements that canonically be
 940 established that regulate the expression of nearby genes where they are found. On the other
 941 hand, we also see the presence of miRNAs, piRNAs and siRNAs in large proportions.
 942 "Others" class represents all those ncRNAs that had less than 3% representation in the
 943 genome, but did have a specific class. **B.** ncRNAs gene prediction was validated by
 944 transcriptional evidence using RNAseq analysis. Our pipeline has a rate of validation
 945 between 82% to 96% for the three evaluated species (*L. braziliensis*, *L. donovani*, *L. major*).
 946 **C.** Comparative expression values of Log2 counts per million (Log2CPM) normalized counts.
 947 Top representative ncRNAs (excluding tRNA and rRNA) according to the number of
 948 predictions had their expression compared in all four strains and with the protein coding
 949 genes available. Each box plot represents a particular isolate for *L. braziliensis* (strain
 950 M2903 and M2904), *L. donovani* and *L. major*.

951

952 **Figure 2. Conservation analysis through *Leishmania* spp.** Flower map indicates the
 953 conservation of all ncRNA in *Leishmania* parasites. Central circle represents the core
 954 RNAome. Accessory ncRNA is represented in the second circle and each petal contains
 955 unique ncRNA found for each evaluated strain. Upset plot corresponds to the top 20
 956 interactions between ncRNAs in all 26 species. Interactions blue represents all the visceral
 957 leishmaniasis (VL) (LdonBH1220, LdonBPK282, LdonAG83, LdonPasteur, LinfJPCM5 and
 958 LinfTR01) related species that exclusively shared ncRNAs clusters. *Viannia* subgenus
 959 (LbraM2903, LbraM2904, LpanPSC1, LpanL13, LperLEM1537, LperPAB4377) shared
 960 ncRNA clusters are in orange. Core ncRNA clusters are represented in purple.

961

Figure 3. Co-expression module-Developmental stages associations. Each row corresponds to a module, column to a developmental stage. Each cell contains the corresponding correlation and p-value (parenthesis). The table is color-coded by correlation according to the color legend. **A.** *L. braziliensis*. **B.** *L. donovani* and **C.** *L. major* Friedlin. PRO = Procyclic promastigote (promastigote in *L. donovani*), META = Metacyclic promastigote and AMA = Amastigote.

Figure 4. GO annotation of ncRNA of Leishmania spp. by module based method performed with GSeq. **A.** *L. braziliensis* **B.** *L. donovani* **C.** *L. major*.

Figure 5. Gene conservation through co-expression modules at different stages of parasitic development and functional annotation. **A.** Coding gene conservation in comparison to *L. braziliensis* and *L. donovani* AMA related modules. **B** and **C.** Coding gene conservation in comparison *L. braziliensis* and *L. major*, PRO and META related modules respectively. **D.** ncRNA conservation in *L. braziliensis* and *L. donovani* AMA related modules comparison. **E** and **F.** ncRNA conservation in *L. braziliensis* and *L. major* PRO and META related modules, respectively. **G.** *L. braziliensis*-*L. donovani* amastigote (AMA) comparison ncRNA functional annotation. **H.** *L. braziliensis*-*L. major* metacyclic promastigote (META) comparison. **I.** *L. braziliensis*-*L. major* procyclic promastigote (PRO) comparison.

Figure 6. Differentially expressed coding and non-coding genes through developmental stages. Coding and ncRNAs expression pairwise comparison through AMA, META and PRO developmental stages. **A.** *L. braziliensis* M2903. In top left overexpression of ncRNAs, bottom left coding genes overexpressed for each comparison. In the right classes of ncRNAs overexpressed in each developmental stage comparison. **B.** *L. donovani* BPK282. **C.** *L. major* Friedlin.

988

989 **Figure 7. Module overrepresentation of differentially expressed genes and ncRNAs**
 990 **through developmental stages in *Leishmania* parasites.** Circos plot showing the gene
 991 expression of coding genes and ncRNAs in relation to the corresponding developmental
 992 stage in *Leishmania* parasites and their module membership. A. Representation for *L.*
 993 *braziliensis* M2903. B. Representation for *L. donovani* BPK282. C. Representation for *L.*
 994 *major* Friedlin.

995

996 **SUPPORTING INFORMATION CAPTIONS**

997 **Supplementary Table 1. non-coding RNA prediction and GC content compared vs**
 998 **genome sequence**

999 **Supplementary Table 2. Accession for LbraM2903, LdonBPK282 and LmajFriedlin**
 1000 **RNA-seq data**

1001 **Supplementary Table 3. Number of all class of non coding RNAs found in genomes of**
 1002 ***Leishmania* spp**

1003 **Supplementary Table 4. Non-coding RNAs transcriptional evidence.**

1004 **Supplementary Table 5. Means of CPM normalized reads values**

1005 **Supplementary Figure 1. ncRNAs classes shared through different *Leishmania***
 1006 **parasites.** Multiclass are all ncRNAs clusters that are composed for more than one different
 1007 class of ncRNAs, for example miRNAs and piRNAs at the same time.

1008 **Supplementary Figure 2. Phylogenetic analysis and comparative with**
 1009 **presence/absence clustering.**

1010 **Supplementary Data 1. ncRNAs GO annotation *Leishmania braziliensis***

1011 **Supplementary Data 2. ncRNAs GO annotation *Leishmania donovani***

1012 **Supplementary Data 3. ncRNAs GO annotation *Leishmania major***

1013 **Supplementary Data 4. Orthologs protein coding genes shared by module**

- 1014 **Supplementary Data 5. Orthologs ncRNAs shared by module**
- 1015 **Supplementary Data 6. ncRNAs GO annotation shared in AMA developmental stage**
- 1016 **Supplementary Data 7. ncRNAs GO annotation shared in META developmental stage**
- 1017 **Supplementary Data 8. ncRNAs GO annotation shared in PRO developmental stage**
- 1018 **Supplementary Data 9. Differentially expressed protein coding genes and ncRNAs in**
- 1019 ***Leishmania braziliensis***
- 1020 **Supplementary Data 10. Differentially expressed protein coding genes and ncRNAs in**
- 1021 ***Leishmania donovani***
- 1022 **Supplementary Data 11. Differentially expressed protein coding genes and ncRNAs in**
- 1023 ***Leishmania major***

A

B

C

A

LbraM2903

B

LdonBPK282

C

LmajFriedlin

Biological Process

- Biosynthetic process
- Catabolic process
- Cellular metabolic process
- Establishment of localization
- Macromolecule metabolic process
- Nitrogen compound metabolic process
- Primary metabolic process
- Protein folding
- Small molecule metabolic process

Molecular Function

- Hydrolase activity
- Protein binding
- Structural component of ribosome
- Transmembrane transporter activity

Cellular component

- Cell part
- DNA packing binding
- Intracellular organelle
- Membrane part
- Non-membrane-bounded organelle
- Organelle
- Protein-DNA complex
- Proteasome complex
- Ribonucleoprotein complex

Amastigote (AMA)

Metacyclic (META)

Promastigote (PRO)

bioRxiv preprint doi: <https://doi.org/10.1101/2021.05.17.444077>; this version posted May 18, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

A

***Leishmania braziliensis* M2903**

B

***Leishmania donovani* BPK282**

C

***Leishmania major* Friedlin**

A *Leishmania braziliensis*

M2903

Coding genes

non-coding RNAs

bioRxiv preprint doi: <https://doi.org/10.1101/2021.05.17.444077>; this version posted May 18, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

B *Leishmania dnonovani* BPK282

Coding genes

non-coding RNAs

C *Leishmania major* Friedlin

Coding genes

non-coding RNAs