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 – CM^2B^2, 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)

ABSTRACT

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

AUTHOR SUMMARY

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

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

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

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

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

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

RESULTS

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

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

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

In order to evaluate the transcriptional evidence for predicted ncRNAs, we used publicly available RNA-seq data (Supplementary Table 2) of four strains from three different 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). After applying an expression cut-off of 0.5 CPM to be considered as a bona fide expressed transcript, we were able to obtain transcriptional evidence for the majority of the predicted ncRNAs. Our analysis revealed that the percentage of expressed ncRNAs range between 83% for *L. donovani* to 95% in *L. braziliensis* (Figure 1B, Supplementary Table 4). Further details of ncRNAs repertoire for each strain are described in Supplementary Table 1 and Supplementary Table 3.

After filtering all normalized values of the top representative ncRNAs, we compared their expression values (Log_2 Counts Per Million) within them and with the protein coding genes for each strain (Figure 1C). As expected, we identified a mean value of CPM for ncRNAs 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.

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-regulatry 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

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

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

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

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

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

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

Module	*L. braziliensis*		*L. donovani*		*L. major*	
	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	-	-

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
L. braziliensis	2834	181	M3 M4 M9	889	44	M3 M4	1081	33	M1 M3 M4 M8 M9 M10

L. donovani	1693	24	M3 M4 M5 M8 M11 M15 M17 M18	-	-		1216	961	M1 M3 M4 M5 M7 M11 M15 M16
L. major	-	-		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

Functional associations between ncRNAs and coding genes

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

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

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

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

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

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

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

To assess the variation in the expression patterns of ncRNAs and coding genes, and their relationship with previous findings across the three different developmental stages, we compared transcriptome experiments performed from PRO, META and AMA stages in *L. 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 obtainedd 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

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

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

correlated with the expression of surface -proteophosphoglycan ppg3-coding gene, suggesiting a potential role in pathogensis for these ncRNAs Similarly, we found that the miRNAs LbraM2903_32_716568_716588 (*L. braziliensis*) and LmajFriedlin_32_1478026_1478046 (*L. major*) co-express negatively with cyclophilin, a protein that act as chaperone and have a role in protein folding (52,53). These results denote the importance of different classes of ncRNAs in a variety of physiological processes in *Leishmania* parasites. Other positive and negative associations between conserved 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_2FC	Class	Module	Coding	Log_2FC	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.19.0340*1	-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

				LmjF.09.1340	-5.15	H2BFM	0.35
LbraM2903_723_1367_1388	2.64	Cis-reg	M3	*LbrM.19.0340*	-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_7_144414_144435	3.41	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
				LmjF.09.1340	-5.15	H2BFM	0.35
LbraM2903_32_716568_716588	2.64	miRNA	M3	*LbrM.19.0340*	-1.59	H2BFM	0.04
				LbrM.22.0710	1.89	hypothetical protein	0.03
				LbrM.34.0610	2.74	ppg3	0.03
				LbrM.34.4360	2.11	4E-interacting protein	0.04
				LbrM.33.1900	-1.61	CYPA4	0.05
LmajFriedlin_32_1478026_1478046	1.71	miRNA	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
				LmjF.09.1340	-5.15	H2BFM	0.35
				LmjF.25.0910	-1.90	CYPA	0.35

DISCUSSION

Here we have described for the first time a genome-wide prediction of non-coding RNAs in 16 species from the *Leishmania* genus. Our results on 26 different strains revealed expression patterns of coding and non-coding RNAs highly related to different parasite developmental stages. These findings may suggest a role in the regulation of morphological 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 repertory 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

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

In order to understand the differences in gene expression along distinct developmental stages in *Leishmania* parasites, we selected three representative strains according to their disease type, subgenus and availability of RNA-seq datasets in public databases. In this sense we compared the expression of both coding genes and ncRNAs from *L.* (*Viannia*) *braziliensis*, which causes cutaneous and mucocutaneous leishmaniasis; *L.* (*Leishmania*) *donovani*, which causes visceral leishmaniasis; and *L.* (*L.*) *major*, that causes cutaneous leishmaniasis. In these three strains we found co-expression patterns in both coding and non-coding RNAs which are exclusively related to specific stages of the developmental cycle of these parasites. We used this information to associate each cluster of ncRNAs to putative functions in *Leishmania* spp. different developmental stages (AMA, PRO and META), based 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 Mabille 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,

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

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

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

In summary, our work is the first to identify novel ncRNAs in 26 genome isolates from 16 Leishmania species. Additionally, it revealed a new set of non-coding genes that are 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.

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.

Comparative analysis of ncRNAs in *Leishmania* spp.

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

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

Co-expression network analysis of developmental stages based on coding genes and 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) mesures, by assigning a gene X as highly related to developmental stage Z, if an X^Z GS ≧ 0.70, p ≦ 0.05 and a MM ≧ |0.5|, p ≦0.05.

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

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

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

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

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

The expression values obtained when determining the transcriptional evidence of coding genes and ncRNAs were used for differential expression analysis. DESeq2 (96) was used in order to determine statistically differentially expressed (DE) coding and non-coding RNAs. To consider a gene as DE we used an empirical probability distribution |log2 fold-change| with a threshold > 1.5 and a *p* value < 0.05.

ACKNOWLEDGMENTS

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

REFERENCES

1. Murray HW, Berman JD, Davies CR, Saravia NG. Advances in leishmaniasis [Internet]. Vol. 366, The Lancet. 2005. p. 1561–77. Available from: http://dx.doi.org/10.1016/s0140-6736(05)67629-5

2. Burza S, Croft SL, Boelaert M. Leishmaniasis. Lancet. 2018 Sep 15;392(10151):951–70.

3. Reithinger R, Dujardin J-C, Louzir H, Pirmez C, Alexander B, Brooker S. Cutaneous leishmaniasis. Lancet Infect Dis. 2007 Sep;7(9):581–96.

4. Subramanian A, Sarkar RR. Perspectives on Leishmania Species and Stage-specific Adaptive Mechanisms. Trends Parasitol. 2018 Dec;34(12):1068–81.

5. WHO | Leishmaniasis. 2020 Jan 13 [cited 2020 Feb 26]; Available from: http://www.who.int/leishmaniasis/en/

6. Bates PA. Revising Leishmania's life cycle [Internet]. Vol. 3, Nature Microbiology. 2018. p. 529–30. Available from: http://dx.doi.org/10.1038/s41564-018-0154-2

7. Podinovskaia M, Descoteaux A. Leishmania and the macrophage: a multifaceted interaction. Future Microbiol. 2015;10(1):111–29.

8. Nieto A, Domínguez-Bernal G, Orden JA, De La Fuente R, Madrid-Elena N, Carrión J. Mechanisms of resistance and susceptibility to experimental visceral leishmaniosis: BALB/c mouse versus Syrian hamster model. Vet Res. 2011 Feb 23;42:39.

9. Rodriguez AE, Estévez JO, Nevot MC, Barrios A, Florin-Christensen M. Leishmania [Internet]. Parasitic Protozoa of Farm Animals and Pets. 2018. p. 289–311. Available from: http://dx.doi.org/10.1007/978-3-319-70132-5_13

10. Dostálová A, Volf P. Leishmania development in sand flies: parasite-vector interactions overview. Parasit Vectors. 2012 Dec 3;5:276.

11. Maran SR, de Lemos Padilha Pitta JL, Dos Santos Vasconcelos CR, McDermott SM, Rezende AM, Silvio Moretti N. Epitranscriptome machinery in Trypanosomatids: New players on the table? Mol Microbiol [Internet]. 2021 Jan 29; Available from: http://dx.doi.org/10.1111/mmi.14688

12. Haile S, Papadopoulou B. Developmental regulation of gene expression in trypanosomatid parasitic protozoa. Curr Opin Microbiol. 2007 Dec;10(6):569–77.

13. Shimoni Y, Friedlander G, Hetzroni G, Niv G, Altuvia S, Biham O, et al. Regulation of gene expression by small non-coding RNAs: a quantitative view [Internet]. Vol. 3, Molecular Systems Biology. 2007. p. 138. Available from: http://dx.doi.org/10.1038/msb4100181

14. Balasubramanian D, Vanderpool CK. New developments in post-transcriptional regulation of operons by small RNAs. RNA Biol. 2013 Mar;10(3):337–41.

15. Vaucheret H. Post-transcriptional small RNA pathways in plants: mechanisms and

regulations. Genes Dev. 2006 Apr 1;20(7):759–71.

16. Sun BK, Tsao H. Small RNAs in development and disease [Internet]. Vol. 59, Journal of the American Academy of Dermatology. 2008. p. 725–37. Available from: http://dx.doi.org/10.1016/j.jaad.2008.08.017

17. Romano G, Veneziano D, Acunzo M, Croce CM. Small non-coding RNA and cancer. Carcinogenesis. 2017 May 1;38(5):485–91.

18. Shirahama S, Miki A, Kaburaki T, Akimitsu N. Long Non-coding RNAs Involved in Pathogenic Infection. Front Genet. 2020 May 26;11:454.

19. Zhang P, Wu W, Chen Q, Chen M. Non-Coding RNAs and their Integrated Networks [Internet]. Vol. 16, Journal of Integrative Bioinformatics. 2019. Available from: http://dx.doi.org/10.1515/jib-2019-0027

20. Fu X-D. Non-coding RNA: a new frontier in regulatory biology. Natl Sci Rev. 2014 Jun 1;1(2):190–204.

21. Koerner MV, Pauler FM, Huang R, Barlow DP. The function of non-coding RNAs in genomic imprinting [Internet]. Vol. 136, Development. 2009. p. 1771–83. Available from: http://dx.doi.org/10.1242/dev.030403

22. Romero-Barrios N, Legascue MF, Benhamed M, Ariel F, Crespi M. Splicing regulation by long noncoding RNAs. Nucleic Acids Res. 2018 Mar 16;46(5):2169–84.

23. Amaral PP, Leonardi T, Han N, Viré E, Gascoigne DK, Arias-Carrasco R, et al. Genomic positional conservation identifies topological anchor point RNAs linked to developmental loci. Genome Biol. 2018 Mar 15;19(1):32.

24. Holley CL, Topkara VK. An introduction to small non-coding RNAs: miRNA and snoRNA. Cardiovasc Drugs Ther. 2011 Apr;25(2):151–9.

25. Chacko N, Lin X. Non-coding RNAs in the development and pathogenesis of eukaryotic microbes. Appl Microbiol Biotechnol. 2013 Sep;97(18):7989–97.

26. Ivens AC, Peacock CS, Worthey EA, Murphy L, Aggarwal G, Berriman M, et al. The genome of the kinetoplastid parasite, Leishmania major. Science. 2005 Jul 15;309(5733):436–42.

27. Peacock CS, Seeger K, Harris D, Murphy L, Ruiz JC, Quail MA, et al. Comparative genomic analysis of three Leishmania species that cause diverse human disease. Nat Genet. 2007 Jul;39(7):839–47.

28. Rogers MB, Hilley JD, Dickens NJ, Wilkes J, Bates PA, Depledge DP, et al. Chromosome and gene copy number variation allow major structural change between species and strains of Leishmania. Genome Res. 2011 Dec;21(12):2129–42.

29. Downing T, Imamura H, Decuypere S, Clark TG, Coombs GH, Cotton JA, et al. Whole genome sequencing of multiple Leishmania donovani clinical isolates provides insights into population structure and mechanisms of drug resistance. Genome Res. 2011 Dec;21(12):2143–56.

30. Real F, Vidal RO, Carazzolle MF, Mondego JMC, Costa GGL, Herai RH, et al. The genome sequence of Leishmania (Leishmania) amazonensis: functional annotation and

extended analysis of gene models. DNA Res. 2013 Dec;20(6):567–81.

31. Zhang W-W, Matlashewski G. Screening Leishmania donovani-specific genes required for visceral infection. Mol Microbiol. 2010 Jul;77(2):505–17.

32. Camacho E, Rastrojo A, Sanchiz Á, la Fuente SG, Aguado B, Requena JM. Leishmania Mitochondrial Genomes: Maxicircle Structure and Heterogeneity of Minicircles [Internet]. Vol. 10, Genes. 2019. p. 758. Available from: http://dx.doi.org/10.3390/genes10100758

33. Inbar E, Keith Hughitt V, Dillon LAL, Ghosh K, El-Sayed NM, Sacks DL. The Transcriptome of Leishmania major Developmental Stages in Their Natural Sand Fly Vector [Internet]. Vol. 8, mBio. 2017. Available from: http://dx.doi.org/10.1128/mbio.00029-17

34. Butenko A, Kostygov AY, Sádlová J, Kleschenko Y, Bečvář T, Podešvová L, et al. Comparative genomics of Leishmania (Mundinia). BMC Genomics. 2019 Oct 11;20(1):726.

35. Kumar A, Pandey SC, Samant M. DNA-based microarray studies in visceral leishmaniasis: identification of biomarkers for diagnostic, prognostic and drug target for treatment. Acta Trop. 2020 May 8;208:105512.

36. Dumas C, Chow C, Müller M, Papadopoulou B. A novel class of developmentally regulated noncoding RNAs in Leishmania. Eukaryot Cell. 2006 Dec;5(12):2033–46.

37. Atayde VD, Shi H, Franklin JB, Carriero N, Notton T, Lye L-F, et al. The structure and repertoire of small interfering RNAs in Leishmania (Viannia) braziliensis reveal diversification in the trypanosomatid RNAi pathway. Mol Microbiol. 2013 Feb;87(3):580–93.

38. Sahoo G, Ansari M, Dikhit M, Gupta N, Rana S, Das P. Computational Identification of microRNA-like Elements in Leishmania major [Internet]. Vol. 2, MicroRNA. 2014. p. 225–30. Available from: http://dx.doi.org/10.2174/2211536602666131203232422

39. Lambertz U, Oviedo Ovando ME, Vasconcelos EJR, Unrau PJ, Myler PJ, Reiner NE. Small RNAs derived from tRNAs and rRNAs are highly enriched in exosomes from both old and new world Leishmania providing evidence for conserved exosomal RNA Packaging. BMC Genomics. 2015 Mar 5;16:151.

40. Eliaz D, Doniger T, Tkacz ID, Biswas VK, Gupta SK, Kolev NG, et al. Genome-wide analysis of small nucleolar RNAs of Leishmania major reveals a rich repertoire of RNAs involved in modification and processing of rRNA. RNA Biol. 2015 May 13;12(11):1222–55.

41. Castro FF, Ruy PC, Zeviani KN, Santos RF, Toledo JS, Cruz AK. Evidence of putative non-coding RNAs from Leishmania untranslated regions [Internet]. Vol. 214, Molecular and Biochemical Parasitology. 2017. p. 69–74. Available from: http://dx.doi.org/10.1016/j.molbiopara.2017.04.002

42. Aoki JI, Muxel SM, Zampieri RA, Laranjeira-Silva MF, Müller KE, Nerland AH, et al. RNA-seq transcriptional profiling of Leishmania amazonensis reveals an arginase-dependent gene expression regulation. PLoS Negl Trop Dis. 2017 Oct;11(10):e0006026.

43. Torres F, Arias-Carrasco R, Caris-Maldonado JC, Barral A, Maracaja-Coutinho V, De Queiroz ATL. LeishDB: a database of coding gene annotation and non-coding RNAs in Leishmania braziliensis. Database [Internet]. 2017 Jan 1;2017. Available from: http://dx.doi.org/10.1093/database/bax047

44. Ruy PDC, De Cássia Ruy P, Monteiro-Teles NM, Magalhães RDM, Freitas-Castro F, Dias L, et al. Comparative transcriptomics in Leishmania braziliensis: disclosing differential gene expression of coding and putative noncoding RNAs across developmental stages [Internet]. Vol. 16, RNA Biology. 2019. p. 639–60. Available from: http://dx.doi.org/10.1080/15476286.2019.1574161

45. Guegan F, Bento F, Neves D, Sequeira M, Notredame C, Figueiredo LM. A long non-coding RNA controls parasite differentiation in African trypanosomes [Internet]. Available from: http://dx.doi.org/10.1101/2020.05.03.074625

46. Paschoal AR, Maracaja-Coutinho V, Setubal JC, Simões ZLP, Verjovski-Almeida S, Durham AM. Non-coding transcription characterization and annotation [Internet]. Vol. 9, RNA Biology. 2012. p. 274–82. Available from: http://dx.doi.org/10.4161/rna.19352

47. Kalvari I, Nawrocki EP, Argasinska J, Quinones-Olvera N, Finn RD, Bateman A, et al. Non-Coding RNA Analysis Using the Rfam Database. Curr Protoc Bioinformatics. 2018 Jun;62(1):e51.

48. Bi K, Chen Y, Zhao S, Kuang Y, John Wu C-H. Current Visceral Leishmaniasis Research: A Research Review to Inspire Future Study. Biomed Res Int. 2018 Jul 10;2018:9872095.

49. Al-Salem W, Herricks JR, Hotez PJ. A review of visceral leishmaniasis during the conflict in South Sudan and the consequences for East African countries. Parasit Vectors. 2016 Aug 22;9:460.

50. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis [Internet]. Vol. 9, BMC Bioinformatics. 2008. Available from: http://dx.doi.org/10.1186/1471-2105-9-559

51. Young MD, Wakefield MJ, Smyth GK, Oshlack A. Gene ontology analysis for RNA-seq: accounting for selection bias. Genome Biol. 2010 Feb 4;11(2):R14.

52. Kofron JL, Kuzmic P, Kishore V, Colón-Bonilla E, Rich DH. Determination of kinetic constants for peptidyl prolyl cis-trans isomerases by an improved spectrophotometric assay. Biochemistry. 1991 Jun 25;30(25):6127–34.

53. Barik S. Immunophilins: for the love of proteins. Cell Mol Life Sci. 2006 Dec;63(24):2889–900.

54. Michaeli S. Non-coding RNA and the complex regulation of the trypanosome life cycle. Curr Opin Microbiol. 2014 Aug;20:146–52.

55. Ibáñez-Ventoso C, Vora M, Driscoll M. Sequence relationships among C. elegans, D. melanogaster and human microRNAs highlight the extensive conservation of microRNAs in biology. PLoS One. 2008 Jul 30;3(7):e2818.

56. Jones-Rhoades MW. Conservation and divergence in plant microRNAs. Plant Mol Biol.

2012 Sep;80(1):3–16.

57. Tsagakis I, Douka K, Birds I, Aspden JL. Long non-coding RNAs in development and disease: Conservation to mechanisms. J Pathol. 2020 Mar 16;8:310433.

58. Sanz-Carbonell A, Marques MC, Martinez G, Gomez G. Dynamic architecture and regulatory implications of the miRNA network underlying the response to stress in melon. RNA Biol. 2020 Feb;17(2):292–308.

59. Barakat A, Wall PK, Diloreto S, Depamphilis CW, Carlson JE. Conservation and divergence of microRNAs in Populus. BMC Genomics. 2007 Dec 31;8:481.

60. Doniger T, Katz R, Wachtel C, Michaeli S, Unger R. A comparative genome-wide study of ncRNAs in trypanosomatids. BMC Genomics. 2010 Nov 4;11:615.

61. Walter Costa MB, Höner Zu Siederdissen C, Dunjić M, Stadler PF, Nowick K. SSS-test: a novel test for detecting positive selection on RNA secondary structure. BMC Bioinformatics. 2019 Mar 21;20(1):151.

62. Cai P, Piao X, Hao L, Liu S, Hou N, Wang H, et al. A deep analysis of the small non-coding RNA population in Schistosoma japonicum eggs. PLoS One. 2013 May 14;8(5):e64003.

63. Mourier T, Carret C, Kyes S, Christodoulou Z, Gardner PP, Jeffares DC, et al. Genome-wide discovery and verification of novel structured RNAs in Plasmodium falciparum. Genome Res. 2008 Feb;18(2):281–92.

64. Liao Q, Shen J, Liu J, Sun X, Zhao G, Chang Y, et al. Genome-wide identification and functional annotation of Plasmodium falciparum long noncoding RNAs from RNA-seq data. Parasitol Res. 2014 Apr;113(4):1269–81.

65. Liao Q, Liu C, Yuan X, Kang S, Miao R, Xiao H, et al. Large-scale prediction of long non-coding RNA functions in a coding–non-coding gene co-expression network. Nucleic Acids Res. 2011 May 1;39(9):3864–78.

66. Dacher M, Tachiwana H, Horikoshi N, Kujirai T, Taguchi H, Kimura H, et al. Incorporation and influence of Leishmania histone H3 in chromatin. Nucleic Acids Res. 2019 Dec 16;47(22):11637–48.

67. Galanti N, Galindo M, Sabaj V, Espinoza I, Toro GC. Histone genes in trypanosomatids. Parasitol Today. 1998 Feb;14(2):64–70.

68. Soto M, Requena JM, Quijada L, Perez MJ, Nieto CG, Guzman F, et al. Antigenicity of the Leishmania infantum histones H2B and H4 during canine viscerocutaneous leishmaniasis. Clin Exp Immunol. 1999 Feb;115(2):342–9.

69. Singh R, Kumar D, Duncan RC, Nakhasi HL, Salotra P. Overexpression of histone H2A modulates drug susceptibility in Leishmania parasites. Int J Antimicrob Agents. 2010 Jul;36(1):50–7.

70. Meddeb-Garnaoui A, Toumi A, Ghelis H, Mahjoub M, Louzir H, Chenik M. Cellular and humoral responses induced by Leishmania histone H2B and its divergent and conserved parts in cutaneous and visceral leishmaniasis patients, respectively. Vaccine. 2010 Feb 17;28(7):1881–6.

71. Carmelo E, Zurita AI, Martinez E, Valladares B. The sera from individuals suffering from cutaneous leishmaniasis due to Leishmania brazilensis present antibodies against parasitic conserved proteins, but not their human counterparts. Parasite. 2006 Sep;13(3):231–6.

72. Zinoviev A, Léger M, Wagner G, Shapira M. A novel 4E-interacting protein in Leishmania is involved in stage-specific translation pathways. Nucleic Acids Res. 2011 Oct;39(19):8404–15.

73. Mabille D, Cardoso Santos C, Hendrickx R, Claes M, Takac P, Clayton C, et al. 4E Interacting Protein as a Potential Novel Drug Target for Nucleoside Analogues in Trypanosoma Brucei. Microorganisms. 2021 Apr 13;9(4):826.

74. Leifso K, Cohen-Freue G, Dogra N, Murray A, McMaster WR. Genomic and proteomic expression analysis of Leishmania promastigote and amastigote life stages: the Leishmania genome is constitutively expressed. Mol Biochem Parasitol. 2007 Mar;152(1):35–46.

75. Patino LH, Muskus C, Ramírez JD. Transcriptional responses of Leishmania (Leishmania) amazonensis in the presence of trivalent sodium stibogluconate. Parasit Vectors. 2019 Jul 12;12(1):348.

76. Qian Z, Yue W, Li B. [Progress of CypA and lung cancer-related research]. Zhongguo Fei Ai Za Zhi. 2010 Aug;13(8):827–31.

77. Liao Y, Luo D, Peng K, Zeng Y. Cyclophilin A: a key player for etiological agent infection. Appl Microbiol Biotechnol. 2021 Feb;105(4):1365–77.

78. Pellé R, McOdimba F, Chuma F, Wasawo D, Pearson TW, Murphy NB. The African trypanosome cyclophilin A homologue contains unusual conserved central and N-terminal domains and is developmentally regulated. Gene. 2002 May 15;290(1-2):181–91.

79. do Monte-Neto RL, Coelho AC, Raymond F, Légaré D, Corbeil J, Melo MN, et al. Gene expression profiling and molecular characterization of antimony resistance in Leishmania amazonensis. PLoS Negl Trop Dis. 2011 May;5(5):e1167.

80. Sayers EW, Beck J, Brister JR, Bolton EE, Canese K, Comeau DC, et al. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 2020 Jan 8;48(D1):D9–16.

81. Aslett M, Aurrecoechea C, Berriman M, Brestelli J, Brunk BP, Carrington M, et al. TriTrypDB: a functional genomic resource for the Trypanosomatidae. Nucleic Acids Res. 2010 Jan;38(Database issue):D457–62.

82. Leinonen R, Sugawara H, Shumway M, on behalf of the International Nucleotide Sequence Database Collaboration. The Sequence Read Archive [Internet]. Vol. 39, Nucleic Acids Research. 2011. p. D19–21. Available from: http://dx.doi.org/10.1093/nar/gkq1019

83. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012 Mar 4;9(4):357–9.

84. Arias-Carrasco R, Vásquez-Morán Y, Nakaya HI, Maracaja-Coutinho V.

StructRNAfinder: an automated pipeline and web server for RNA families prediction. BMC Bioinformatics. 2018 Feb 17;19(1):55.

85. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features [Internet]. Vol. 26, Bioinformatics. 2010. p. 841–2. Available from: http://dx.doi.org/10.1093/bioinformatics/btq033

86. Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data [Internet]. [cited 2020 Feb 26]. Available from: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

87. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014 Aug 1;30(15):2114–20.

88. Anders S, Pyl PT, Huber W. HTSeq--a Python framework to work with high-throughput sequencing data. Bioinformatics. 2015 Jan 15;31(2):166–9.

89. Ramos TAR, Maracaja-Coutinho V, Ortega JM, do Rêgo TG. CORAZON: a web server for data normalization and unsupervised clustering based on expression profiles. BMC Res Notes. 2020 Jul 14;13(1):338.

90. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics. 2012 Dec 1;28(23):3150–2.

91. Lechner M, Findeiss S, Steiner L, Marz M, Stadler PF, Prohaska SJ. Proteinortho: detection of (co-)orthologs in large-scale analysis. BMC Bioinformatics. 2011 Apr 28;12:124.

92. Katoh K, Misawa K, Kuma K-I, Miyata T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. 2002 Jul 15;30(14):3059–66.

93. Castresana J. Selection of Conserved Blocks from Multiple Alignments for Their Use in Phylogenetic Analysis [Internet]. Vol. 17, Molecular Biology and Evolution. 2000. p. 540–52. Available from: http://dx.doi.org/10.1093/oxfordjournals.molbev.a026334

94. Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol. 2015 Jan;32(1):268–74.

95. Xu L, Dong Z, Fang L, Luo Y, Wei Z, Guo H, et al. OrthoVenn2: a web server for whole-genome comparison and annotation of orthologous clusters across multiple species. Nucleic Acids Res. 2019 Jul 2;47(W1):W52–8.

96. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2 [Internet]. Vol. 15, Genome Biology. 2014. Available from: http://dx.doi.org/10.1186/s13059-014-0550-8

FIGURE CAPTIONS

Figure 1. ncRNAs repertoire and transcriptional evidence in *Leishmania* parasites. A. The frequency of all predicted ncRNAs for each strain of *Leishmania* spp. evaluated. Those ncRNAs representing less than 0.5% were categorized as others. We can emphasize the presence mainly of Cis-regulatory RNA, these RNAs are elements that canonically be established that regulate the expression of nearby genes where they are found. On the other hand, we also see the presence of miRNAs, piRNAs and siRNAS in large proportions. "Others"' class represents all those ncRNAs that had less than 3% representation in the genome, but did have a specific class. **B.** ncRNAs gene prediction was validated by transcriptional evidence using RNAseq analysis. Our pipeline has a rate of validation between 82% to 96% for the three evaluated species (*L. braziliensis, L. donovani, L. major*). **C.** Comparative expression values of Log2 counts per million (Log2CPM) normalized counts**.** Top representative ncRNAs (excluding tRNA and rRNA) according to the number of predictions had their expression compared in all four strains and with the protein coding genes available. Each box plot represents a particular isolate for *L. braziliensis* (strain M2903 and M2904), *L. donovani* and *L. major*.

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

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 GOseq. 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*-*Lmajor* 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.

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

SUPPORTING INFORMATION CAPTIONS

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

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

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

Supplementary Table 4. Non-coding RNAs transcriptional evidence.

Supplementary Table 5. Means of CPM normalized reads values

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

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

Supplementary Data 1. ncRNAs GO annotation *Leishmania braziliensis*

Supplementary Data 2. ncRNAs GO annotation *Leishmania donovani*

Supplementary Data 3. ncRNAs GO annotation *Leishmania major*

Supplementary Data 4. Orthologs protein coding genes shared by module

Supplementary Data 5. Orthologs ncRNAs shared by module

Supplementary Data 6. ncRNAs GO annotation shared in AMA developmental stage

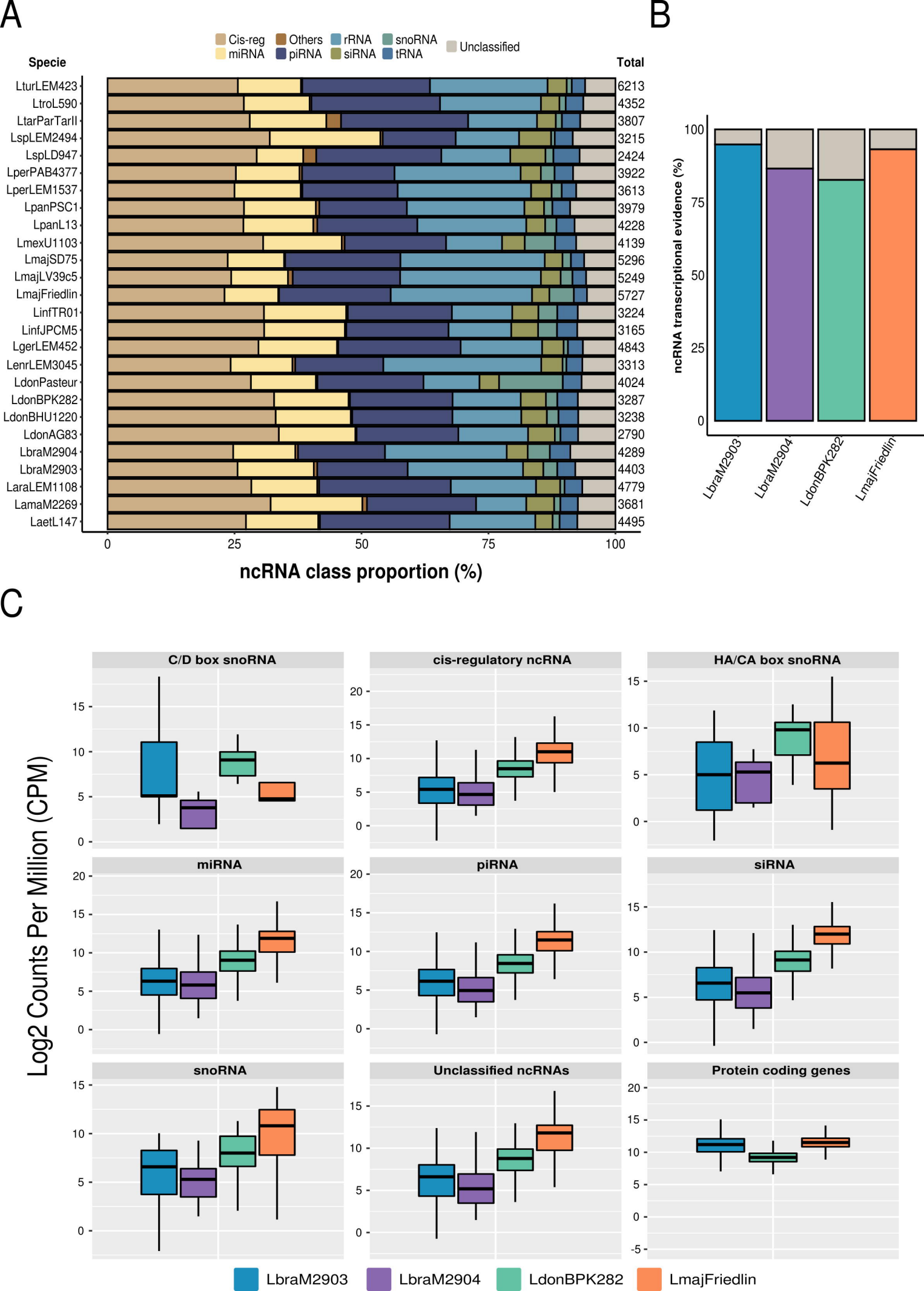
Supplementary Data 7. ncRNAs GO annotation shared in META developmental stage

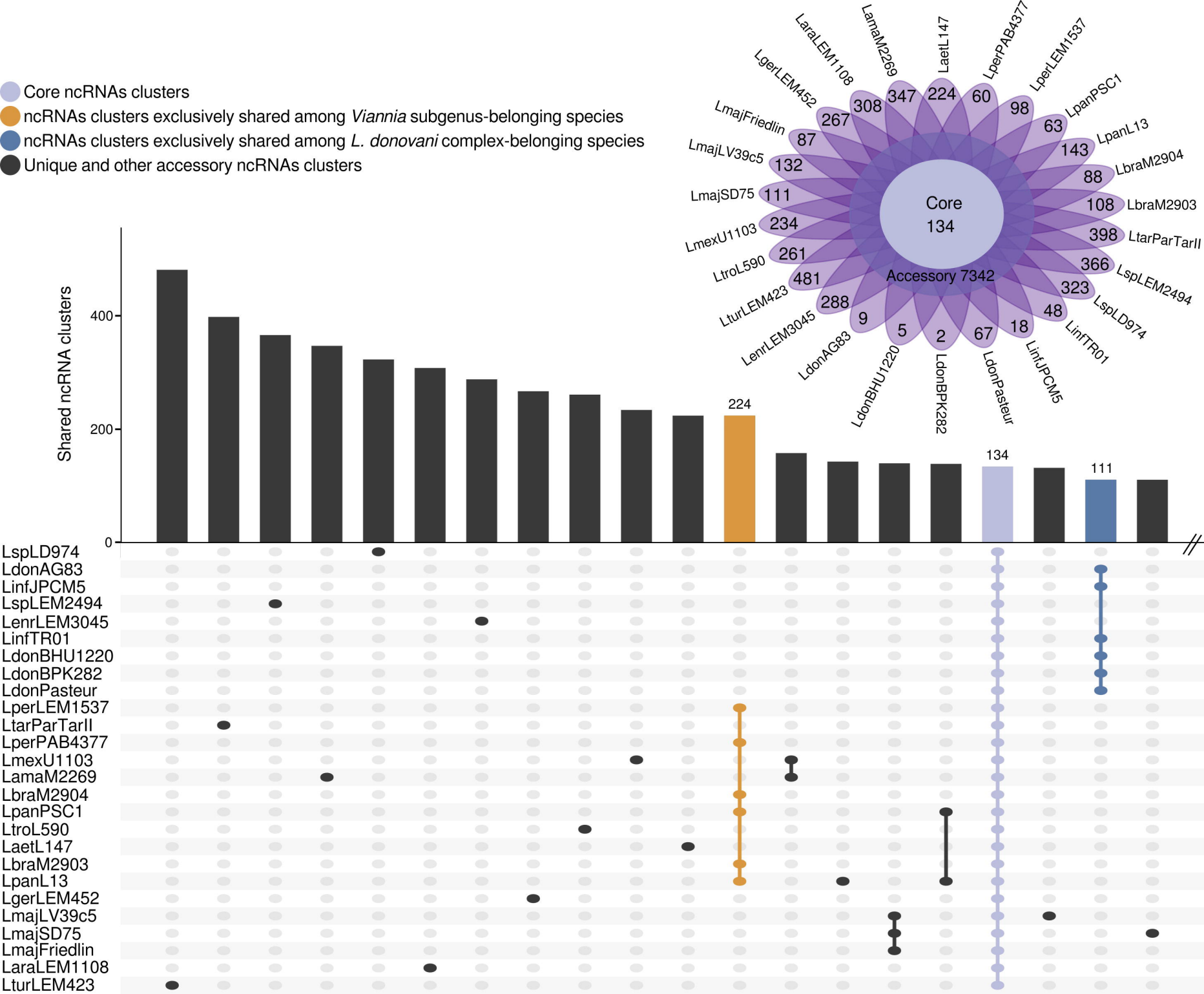
Supplementary Data 8. ncRNAs GO annotation shared in PRO developmental stage

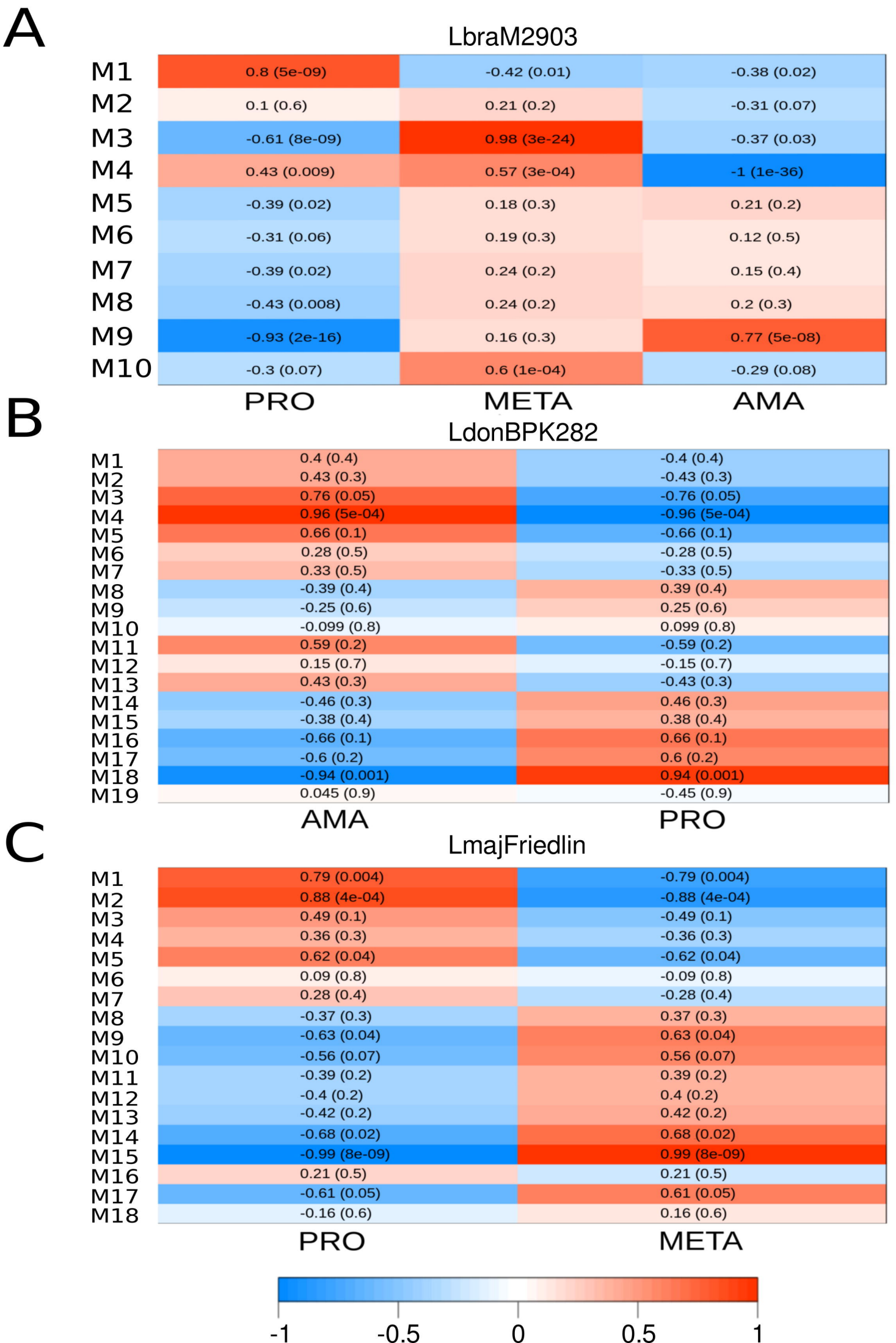
Supplementary Data 9. Differentially expressed protein coding genes and ncRNAs in *Leishmania braziliensis*

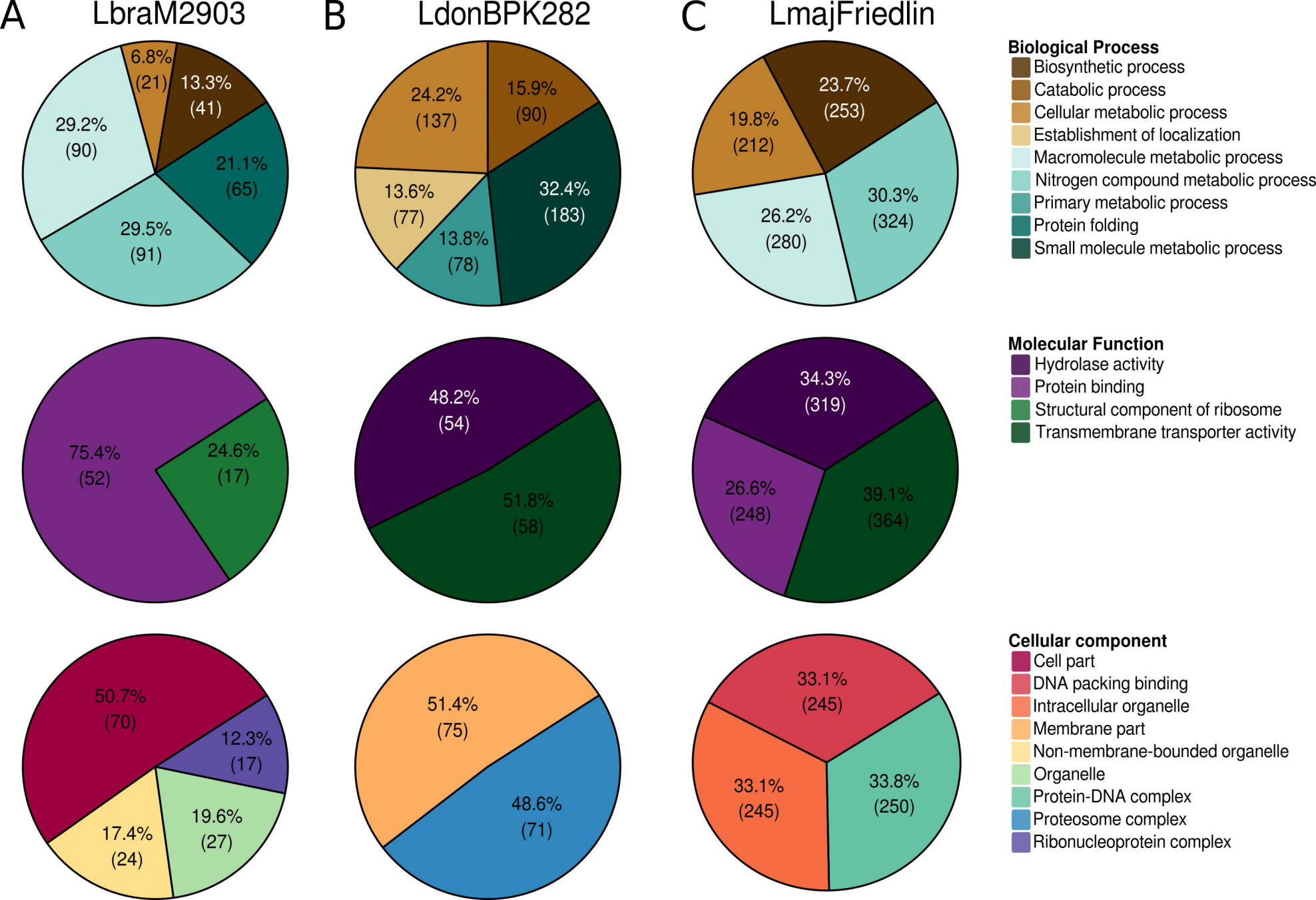
Supplementary Data 10. Differentially expressed protein coding genes and ncRNAs in *Leishmania donovani*

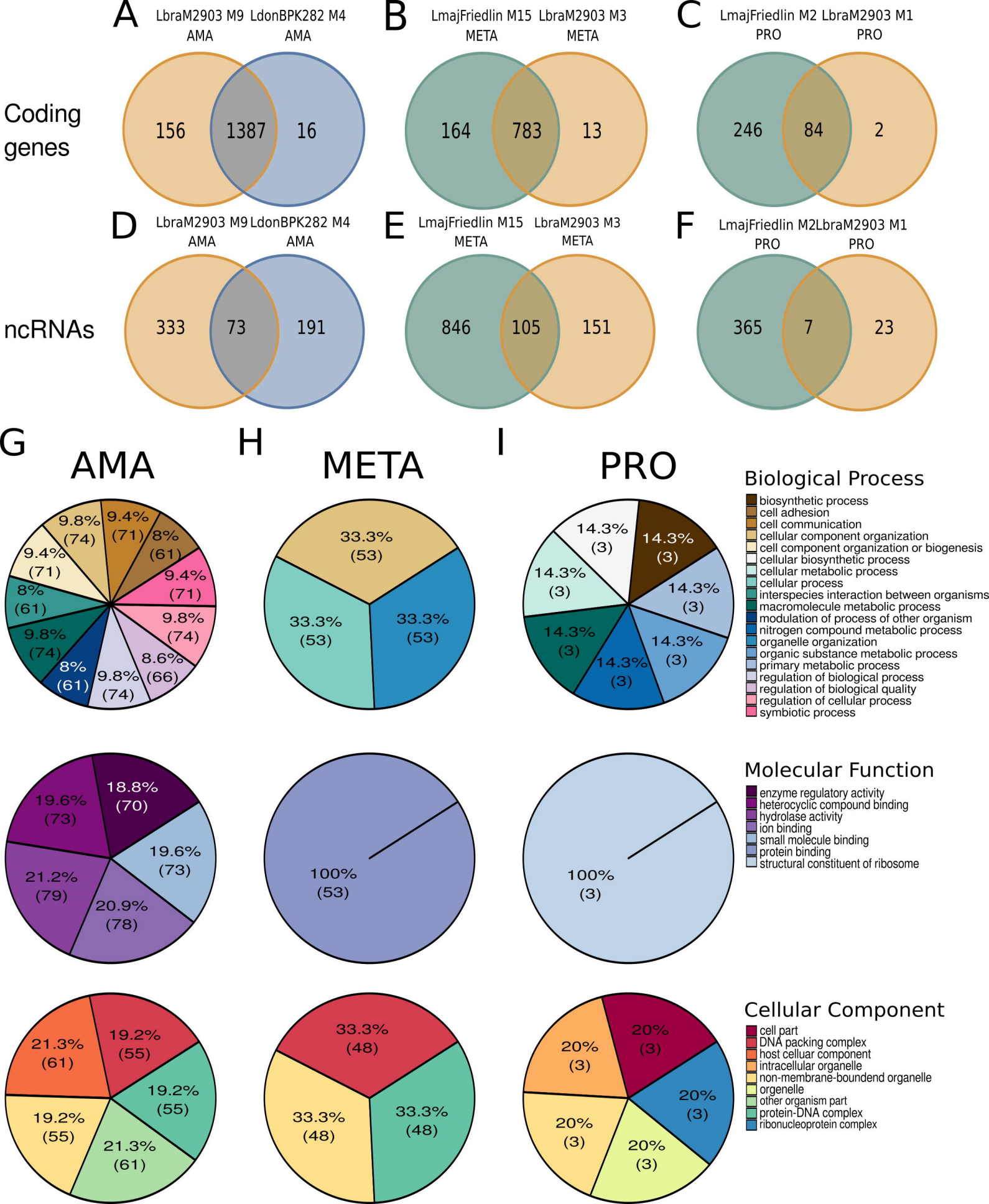
Supplementary Data 11. Differentially expressed protein coding genes and ncRNAs in *Leishmania major*

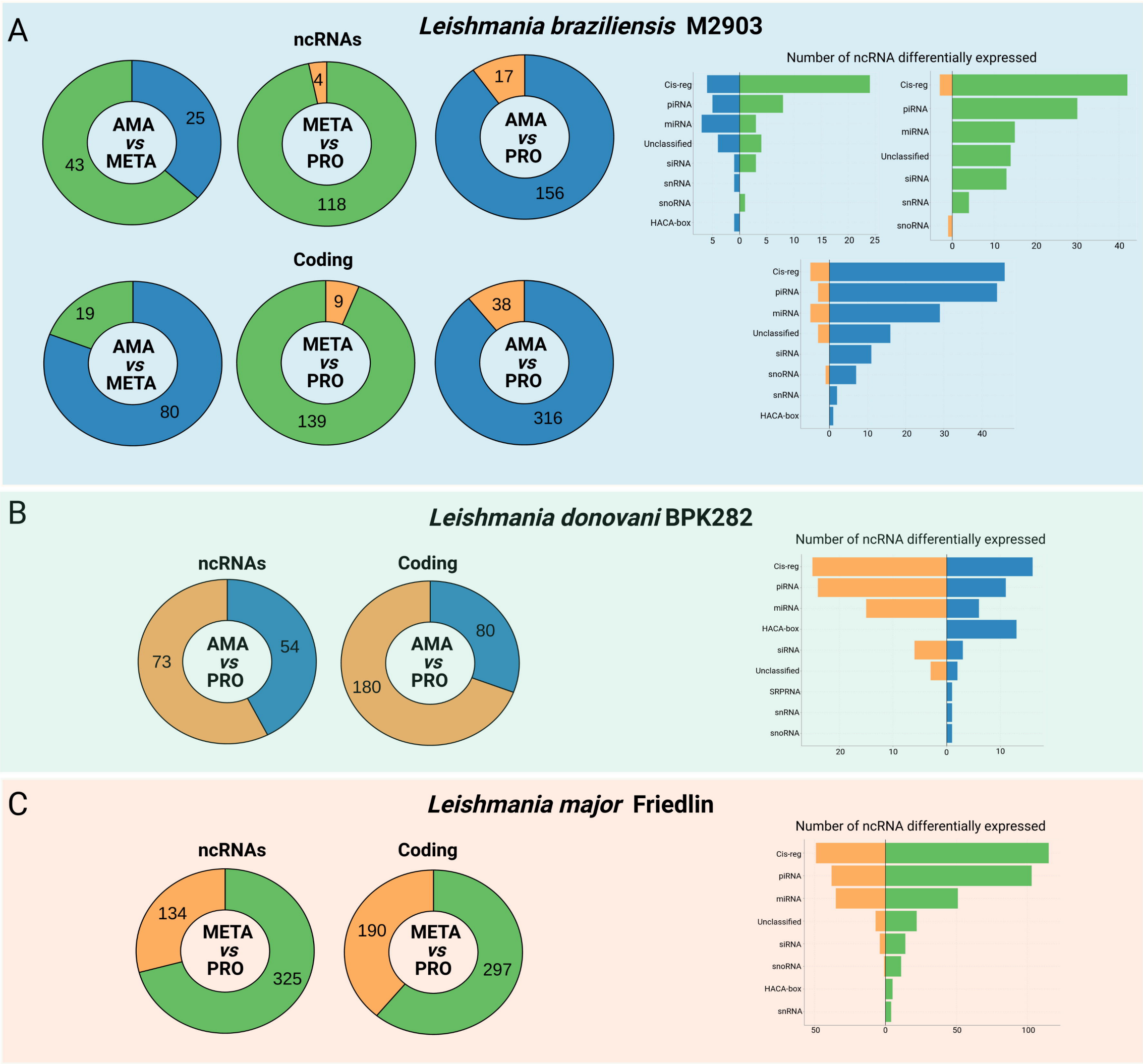

A
Cis-reg
Others
rRNA
snoRNA
Unclassified
miRNA
piRNA
siRNA
tRNA
Specie
Total
LturLEM423 6213
LtroL590 4352
LtarParTarII 3807
LspLEM2494 3215
LspLD947 2424
LperPAB4377 3922
LperLEM1537 3613
LpanPSC1 3979
LpanL13 4228
LmexU1103 4139
LmajSD75 5296
LmajLV39c5 5249
LmajFriedlin 5727
LinfTR01 3224
LinfJPCM5 3165
LgerLEM452 4843
LenrLEM3045 3313
LdonPasteur 4024
LdonBPK282 3287
LdonBHU1220 3238
LdonAG83 2790
LbraM2904 4289
LbraM2903 4403
LaraLEM1108 4779
LamaM2269 3681
LaetL147 4495
0
25
50
75
100
ncRNA class proportion (%)
B
ncRNA transcriptional evidence (%)
100
75
50
25
0
LbraM2903
LbraM2904
LdonBPK282
LmajFriedlin
C
Log2 Counts Per Million (CPM)
C/D box snoRNA
cis-regulatory ncRNA
HA/CA box snoRNA
miRNA
piRNA
siRNA
snoRNA
Unclassified ncRNAs
Protein coding genes
LbraM2903
LbraM2904
LdonBPK282
LmajFriedlin

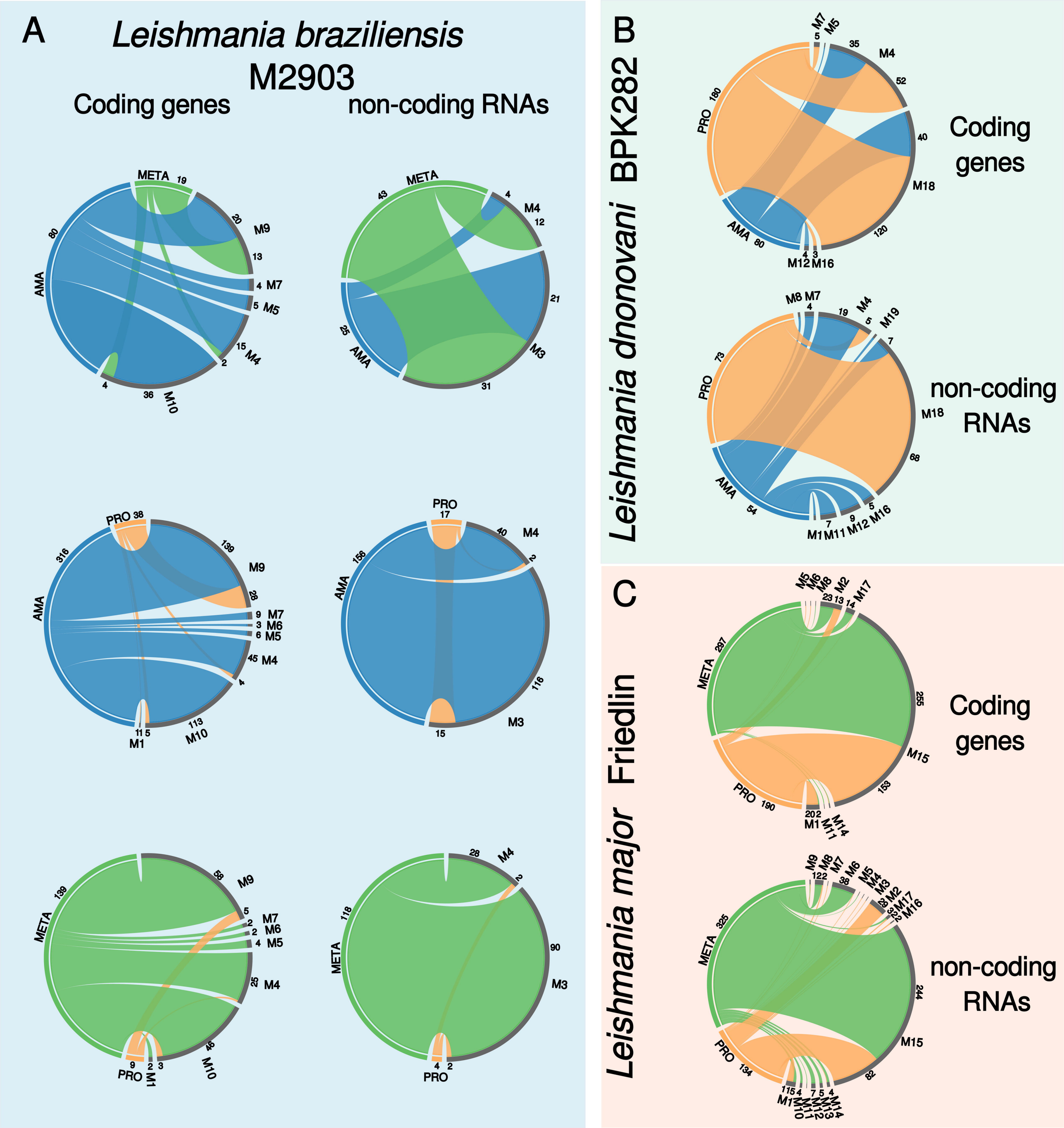

Core ncRNAs clusters
ncRNAs clusters exclusively shared among Viannia subgenus-belonging species
ncRNAs clusters exclusively shared among L. donovani complex-belonging species
Unique and other accessory ncRNAs clusters
Core
134
Accessory 7342
LaraLEM1108
LamaM2269
LaetL147
LperPAB4377
LperLEM1537
LgerLEM452
LpanPSC1
LmajFriedlin
LpanL13
LmajLV39c5
LbraM2904
LmajSD75
LbraM2903
LmexU1103
LtarParTarII
LtroL590
LspLEM2494
LturLEM423
LspLD974
LenrLEM3045
LinfTR01
LdonAG83
LinfJPCM5
LdonBHU1220
LdonPasteur
LdonBPK282
347
224
60
98
63
143
88
108
398
366
323
48
18
67
2
5
9
288
481
261
234
111
132
87
267
308
Shared ncRNA clusters
0
200
400
224
134
111
LspLD974
LdonAG83
LinfJPCM5
LspLEM2494
LenrLEM3045
LinfTR01
LdonBHU1220
LdonBPK282
LdonPasteur
LperLEM1537
LtarParTarII
LperPAB4377
LmexU1103
LamaM2269
LbraM2904
LpanPSC1
LtroL590
LaetL147
LbraM2903
LpanL13
LgerLEM452
LmajLV39c5
LmajSD75
LmajFriedlin
LaraLEM1108
LturLEM423


A
LbraM2903
M1 0.8 (5e-09) -0.42 (0.01) -0.38 (0.02)
M2 0.1 (0.6) 0.21 (0.2) -0.31 (0.07)
M3 -0.61 (8e-09) 0.98 (3e-24) -0.37 (0.03)
M4 0.43 (0.009) 0.57 (3e-04) -1 (1e-36)
M5 -0.39 (0.02) 0.18 (0.3) 0.21 (0.2)
M6 -0.31 (0.06) 0.19 (0.3) 0.12 (0.5)
M7 -0.39 (0.02) 0.24 (0.2) 0.15 (0.4)
M8 -0.43 (0.008) 0.24 (0.2) 0.2 (0.3)
M9 -0.93 (2e-16) 0.16 (0.3) 0.77 (5e-08)
M10 -0.3 (0.07) 0.6 (1e-04) -0.29 (0.08)
PRO META AMA
B
LdonBPK282
M1 0.4 (0.4) -0.4 (0.4)
M2 0.43 (0.3) -0.43 (0.3)
M3 0.76 (0.05) -0.76 (0.05)
M4 0.96 (5e-04) -0.96 (5e-04)
M5 0.66 (0.1) -0.66 (0.1)
M6 0.28 (0.5) -0.28 (0.5)
M7 0.33 (0.5) -0.33 (0.5)
M8 -0.39 (0.4) 0.39 (0.4)
M9 -0.25 (0.6) 0.25 (0.6)
M10 -0.099 (0.8) 0.099 (0.8)
M11 0.59 (0.2) -0.59 (0.2)
M12 0.15 (0.7) -0.15 (0.7)
M13 0.43 (0.3) -0.43 (0.3)
M14 -0.46 (0.3) 0.46 (0.3)
M15 -0.38 (0.4) 0.38 (0.4)
M16 -0.66 (0.1) 0.66 (0.1)
M17 -0.6 (0.2) 0.6 (0.2)
M18 -0.94 (0.001) 0.94 (0.001)
M19 0.045 (0.9) -0.45 (0.9)
AMA PRO
C
LmajFriedlin
M1 0.79 (0.004) -0.79 (0.004)
M2 0.88 (4e-04) -0.88 (4e-04)
M3 0.49 (0.1) -0.49 (0.1)
M4 0.36 (0.3) -0.36 (0.3)
M5 0.62 (0.04) -0.62 (0.04)
M6 0.09 (0.8) -0.09 (0.8)
M7 0.28 (0.4) -0.28 (0.4)
M8 -0.37 (0.3) 0.37 (0.3)
M9 -0.63 (0.04) 0.63 (0.04)
M10 -0.56 (0.07) 0.56 (0.07)
M11 -0.39 (0.2) 0.39 (0.2)
M12 -0.4 (0.2) 0.4 (0.2)
M13 -0.42 (0.2) 0.42 (0.2)
M14 -0.68 (0.02) 0.68 (0.02)
M15 -0.99 (8e-09) 0.99 (8e-09)
M16 0.21 (0.5) 0.21 (0.5)
M17 -0.61 (0.05) 0.61 (0.05)
M18 -0.16 (0.6) 0.16 (0.6)
PRO META
-1 -0.5 0 0.5 1


A
LbraM2903
B
LdonBPK282
C
LmajFriedlin
6.8% (21)
13.3% (41)
29.2% (90)
21.1% (65)
29.5% (91)
24.2% (137)
15.9% (90)
13.6% (77)
32.4% (183)
13.8% (78)
23.7% (253)
19.8% (212)
26.2% (280)
30.3% (324)
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
75.4% (52)
24.6% (17)
48.2% (54)
51.8% (58)
34.3% (319)
26.6% (248)
39.1% (364)
Molecular Function
Hydrolase activity
Protein binding
Structural component of ribosome
Transmembrane transporter activity
50.7% (70)
12.3% (17)
17.4% (24)
19.6% (27)
51.4% (75)
48.6% (71)
33.1% (245)
33.1% (245)
33.8% (250)
Cellular component
Cell part
DNA packing binding
Intracellular organelle
Membrane part
Non-membrane-bounded organelle
Organelle
Protein-DNA complex
Proteosome complex
Ribonucleoprotein complex


A
LbraM2903 M9 AMA
LdonBPK282 M4 AMA
156
1387
16
B
LmajFriedlin M15 META
LbraM2903 M3 META
164
783
13
C
LmajFriedlin M2 PRO
LbraM2903 M1 PRO
246
84
2
Coding genes
D
LbraM2903 M9 AMA
LdonBPK282 M4 AMA
333
73
191
E
LmajFriedlin M15 META
LbraM2903 M3 META
846
105
151
F
LmajFriedlin M2 PRO
LbraM2903 M1 PRO
365
7
23
ncRNAs
G
AMA
9.8% (74)
9.4% (71)
8% (61)
9.4% (71)
9.8% (74)
8.6% (66)
9.8% (74)
8% (61)
9.8% (74)
8% (61)
9.4% (71)
H
META
33.3% (53)
33.3% (53)
33.3% (53)
I
PRO
14.3% (3)
Biological Process
biosynthetic process
cell adhesion
cell communication
cellular component organization
cell component organization or biogenesis
cellular biosynthetic process
cellular metabolic process
cellular process
interspecies interaction between organisms
macromolecule metabolic process
modulation of process of other organism
nitrogen compound metabolic process
organelle organization
organic substance metabolic process
primary metabolic process
regulation of biological process
regulation of biological quality
regulation of cellular process
symbiotic process
18.8% (70)
19.6% (73)
19.6% (73)
21.2% (79)
20.9% (78)
100% (53)
100% (3)
Molecular Function
enzyme regulatory activity
heterocyclic compound binding
hydrolase activity
ion binding
small molecule binding
protein binding
structural constituent of ribosome
21.3% (61)
19.2% (55)
19.2% (55)
19.2% (55)
21.3% (61)
33.3% (48)
33.3% (48)
33.3% (48)
20% (3)
Cellular Component
cell part
DNA packing complex
host celluar component
intracellular organelle
non-membrane-boundend organelle
orgenelle
other organism part
protein-DNA complex
ribonucleoprotein complex


Amastigote (AMA)
Metacyclic (META)
Promastigote (PRO)
A
Leishmania braziliensis M2903
ncRNAs
AMA vs META
25
43
META vs PRO
4
118
AMA vs PRO
17
156
Coding
AMA vs META
19
80
META vs PRO
9
139
AMA vs PRO
38
316
Number of ncRNA differentially expressed
Cis-reg
piRNA
miRNA
Unclassified
siRNA
snRNA
snoRNA
HACA-box
B
Leishmania donovani BPK282
ncRNAs
AMA vs PRO
73
54
Coding
AMA vs PRO
180
80
Number of ncRNA differentially expressed
Cis-reg
piRNA
miRNA
HACA-box
siRNA
Unclassified
SRPRNA
snRNA
snoRNA
C
Leishmania major Friedlin
ncRNAs
META vs PRO
134
325
Coding
META vs PRO
190
297
Number of ncRNA differentially expressed
Cis-reg
piRNA
miRNA
Unclassified
siRNA
snoRNA
HACA-box
snRNA


A
Leishmania braziliensis
M2903
Coding genes
non-coding RNAs
B
Leishmania donovani BPK282
Coding genes
non-coding RNAs
C
Leishmania major Friedlin
Coding genes
non-coding RNAs
META
AMA
PRO
M1
M3
M4
M5
M6
M7
M8
M9
M10
M11
M12
M14
M15
M16
M17
M18
M19