Opinion

Cell PRESS

Sequence-based functional annotation: what if most of the genes are unique to a genome?

Reza Salavati[1,2,3] and Hamed Shateri Najafabadi[1,2]

[1] Institute of Parasitology, McGill University, 21,111 Lakeshore Road, Ste. Anne de Bellevue, Montreal, Quebec H9X3V9, Canada
[2] McGill Center for Bioinformatics, McGill University, Duff Medical Building, 3775 University Street, Montreal, Quebec H3A2B4, Canada
[3] Department of Biochemistry, McGill University, McIntyre Medical Building, 3655 Promenade Sir William Osler, Montreal, Quebec H3G1Y6, Canada

The genomes of trypanosomatids are distantly related to other eukaryotes, with significant numbers of hypothetical or conserved hypothetical trypanosomatid-specific genes, whose functions cannot be determined using homology-dependent annotation methods. Here, we describe homology-independent methods to infer biological functions of genes based solely on their sequences. These approaches are not limited to trypanosomatid genomes and provide grounds for analysis of genomes of *Plasmodium falciparum* and other parasites associated with neglected tropical diseases. A critical evaluation of the current state of annotation of parasitic genomes endorses the need to exploit homology-independent computational methods, which can identify protein functions, potentially including essential genes, and provide a plethora of valuable information on interaction networks and regulatory elements.

Genome annotation of trypanosomatids and its limitations

Trypanosomatid pathogens are responsible for serious human and animal diseases, with a very high mortality rate if untreated. There are no vaccines for these pathogens, the available drugs are toxic with limited effectiveness, and drug resistance is emerging. Although the genome sequences are available for the most prominent trypanosomatids, *Trypanosoma brucei*, *T. cruzi*, *Leishmania major*, *L. infantum* and *L. braziliensis* [1–5], a high percentage of their genes are non-annotated, limiting the available drug targets to the subset of genes whose functions are known or can be inferred from homology. The focus on three species (i.e. *T. brucei*, *T. cruzi* and *L. major* – collectively called TriTryps) has led the EuPath Project Team to launch TriTrypDB (http://TriTrypDB.org) with the aim of providing an integrated genomic and functional database for trypanosomatids. Although this database offers a wealth of resources to query TriTryp genomes, it still lacks a comprehensive functional annotation of their genes in that homology-based genome annotation in trypanosomatids is limited by the poor sequence similarity between the genomes of trypanosomatids and the genomes of other sequenced organisms, particularly eukaryotes such as human, yeast and *Caenorhabditis elegans*, in which gene functions are extensively studied. For example, out of approximately 9100 predicted and validated genes in *T. brucei*, around 4900 have no reliable homologs in the sequenced genomes of non-trypanosomatid organisms (blastp, E-value $\leq 1\times10^{-6}$). Not all the remaining ~4200 genes can be assigned a function, because some only have homologs that are also uncharacterized. In fact, approximately 35% of these conserved genes are annotated just as 'hypothetical'. Currently, only about 3400 *T. brucei* genes have any annotation other than hypothetical (Figure 1).

However, current and developing methods for computational prediction of gene function hold a great promise to facilitate the functional annotation of trypanosomatid genomes. Methods other than homology-based transfer of annotations can help to annotate these genomes (see below). Many methods have emerged recently that can predict the likely functions and interactions of genes independent of the presence of homologs in other organisms. Using these methods in combination with homology-based approaches, it seems very likely that a considerable number of currently unannotated genes can readily be assigned biological functions.

Computational annotation of the genome

In addition to the direct search for characterized homologs of a gene (i.e. through BLAST), other methods have been established by which gene functions can be inferred. Network-based approaches exploit the observation that proteins with related functions usually interact with each

Glossary

Hypothetical protein: a protein whose expression is not supported by any experimental evidences, and whose function is unknown.
ORFeome: the collection of ORFs (open reading frames) in a genome.
Precision: the fraction of objects (e.g. genes) that are predicted to be in a particular category (e.g. a biological process) that actually belong to that category. It is also frequently referred to as Positive Predictive Value.
Sensitivity: here, it means the fraction of genes within a biological process that are correctly assigned to that process by a prediction method.
Specificity: here, it means the fraction of genes outside a biological process that are correctly predicted as not being involved in that process.

Corresponding author: Salavati, R. (reza.salavati@mcgill.ca).

1471-4922/$ – see front matter doi:10.1016/j.pt.2010.02.001 Available online 6 March 2010

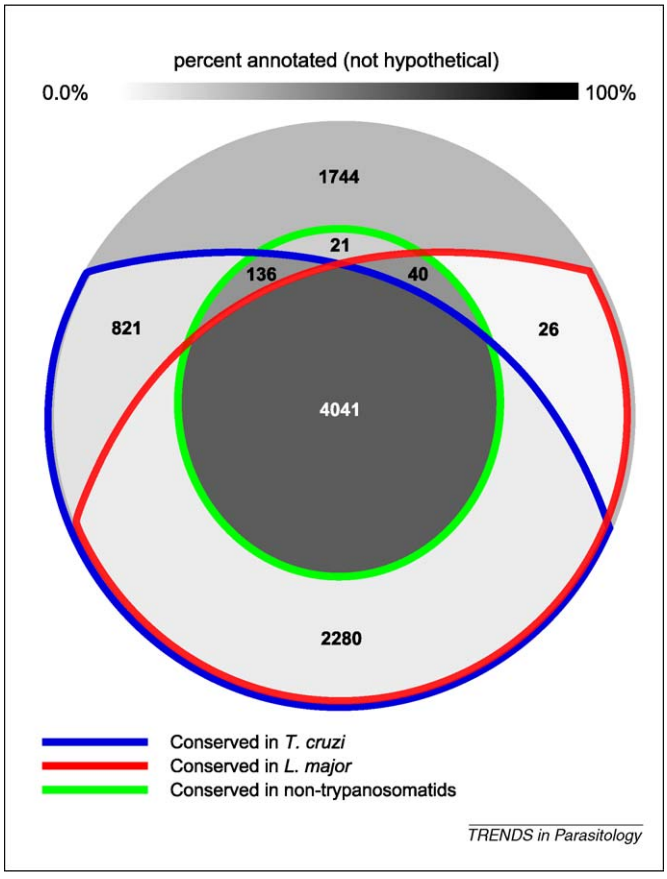

Figure 1. Only a small fraction of trypanosomatid genes currently have functional annotations. In this figure, *Trypanosoma brucei* proteins are compared to *T. cruzi* and *Leishmania major* proteins as well as the proteins of all other organisms with available genome sequences (blastp, E-value $\leq 1\times10^{-6}$). The fraction of *T. brucei* ORFeome that is conserved in *T. cruzi* is bordered by the blue curve. The red curve borders the proteins that are conserved in *L. major*, and the green circle indicates conservation in any of 1019 non-trypanosomatid organisms with available ORFeome sequences on KEGG database. Out of approximately 6300 genes shared among *T. brucei*, *L. major* and *T. cruzi*, around 4000 can also be found in non-trypanosomatid organisms with known genome sequences, of which less than 65% currently have any functional annotation (see the color legend above the figure). The relatively high percentage of annotation of *T. brucei*-specific genes (the top fraction) reflects the large number of variant surface glycoproteins (there are more than 1000 *T. brucei*-specific variant surface glycoprotein genes in the current release of *T. brucei* genome).

other, and thus cluster together in the network of protein–protein interactions. Several different approaches have been used to assign functions based on: (i) protein–protein interactions (reviewed in Ref. [6]); (ii) the clustering of genes according to expression patterns [7] (genes with similar expression patterns have related functions [8–10]); or (iii) the presence of conserved motifs within protein sequences. The combination of these three (i.e. interaction networks, expression patterns and protein motifs) has been shown to be superior to any one of them alone, but interaction networks claim the major share, contributing to approximately 85% of predictions [11]. As the genome-wide interaction network is the most informative indicator of functional linkages between proteins, it is crucial to obtain such a network. In the absence of experimental data, several computational methods have been used to predict protein–protein interactions [12]. Combination of these methods has proved powerful for computational modeling of interaction networks and functional linkages [13–15]. However, many of the prominent current methods rely on the presence of homologs in other species [16–21], limiting their application to only a subset of genes for use in trypanosomatids.

Use of a novel approach based on codon usage for genome annotation

A recent method, called PIC (Probabilistic-Interactome using Codon usage) [22], has been shown to be able to predict functional linkages and/or physical interactions of proteins based on similarity of codon usages of their corresponding genes. Because this method does not rely on cross-species homology, it can be used for detection of linkages between any protein pairs. This method was initially shown to work for *Saccharomyces cerevisiae*, *Plasmodium falciparum* and *Escherichia coli* [22], particularly when combined with other approaches (see below). Later, a large-scale analysis of all sequenced genomes showed that codon usage and gene function are two correlated properties in almost all organisms [23], including trypanosomatids. Based on this observation, an improved algorithm was developed that could directly predict the function of a gene (based on its codon usage). As an example, this algorithm was able to find *T. brucei* genes that are involved in inositol phosphate metabolism with >99% specificity at sensitivities up to 7% [23]. Other examples included ribosome, benzoate degradation via CoA ligation and phosphatidylinositol signaling system. Although this sensitivity on its own is not very exciting, it suggests that the combination of this method with other homology-independent methods can build a powerful classifier, as discussed below.

Use of regulatory elements in genome annotation

Genes in trypanosomatids are transcribed as polycistronic mRNAs, which are further processed via trans-splicing, involving a polypyrimidine tract as the signal for spliced-leader site [24]. This feature can be used for prediction of splice sites and, less confidently, polyadenylation sites from the genomic sequence, giving reasonable estimates for the mature mRNA ends. Regulation of gene expression in trypanosomatids is mainly at the post-transcriptional level by either regulation of mRNA stability or translation [25,26]. However, a few regulatory elements have been identified, all of which are in the 3′ UTR of developmentally regulated genes [27–50]. Some indications suggest that elements in regions other than 3′ UTRs could also play roles in developmental regulation of expression [38], but none has yet been identified.

In a recent study [51], a computational analysis of *T. brucei* genome was conducted to identify statistically reliable function-specific sequence motifs. This study also presented a method to predict gene function based on these potential regulatory elements [51]. Regulatory motifs within 3′ and 5′ UTRs of functionally related genes were predicted using FIRE (finding informative regulatory elements), a method that had previously been designed and applied successfully for finding informative regulatory elements [52]. This resulted in 15 function-specific motifs in 5′ UTRs of *T. brucei* genes and 21 function-specific motifs in their 3′ UTRs, with an overall estimated precision of 75.3% for discovering function-specific 5′ UTR motifs and

84.8% for 3′ UTR motifs [51]. The found regulatory motifs covered a wide range of different pathways from glycolysis to DNA replication. Once experimentally validated, these motifs can provide new insights on the regulatory mechanisms of trypanosomatids and possible developmental regulation of genes.

Although these motifs have not yet been confirmed experimentally, it is shown that a naive Bayesian network can effectively predict many gene functions in *T. brucei* using the pattern of presence or absence of these predicted regulatory motifs [51]. For example, a sensitivity of 20% could be reached at a specificity of ~99% for predicting proteins involved in the inositol phosphate metabolism pathway (precision: 55%). This prompted us to test whether a combination of codon usage [23] (and see above) and regulatory motifs [51] could make a robust gene–function predictor for this particular pathway. It was found that such a combination via a simple naive Bayesian network can achieve up to 50% sensitivity with >60% precision in identification of genes involved in inositol phosphate metabolism (Figure 2). The results of this combination, for genes that are not trypanosomatid-specific, are consistent with results from homology-based mapping of protein–protein interactions, which underpins the method (R. Salavati, unpublished).

Other possibilities for homology-independent annotation of genomes

In the previous section, we explained the possibility of using function-specific regulatory nucleotide motifs for function prediction. A less explored possibility, however, is the use of a similar approach for identifying function-specific 'linear protein motifs'. Proteins with related biological functions are, in many cases, regulated post-translationally via similar peptide patterns; these post-translational modifications are widely used in parasitic cells (see Ref. [53] for a review of post-translational modifications in *Plasmodium*). Proteins with similar molecular functions can also share common peptide patterns that represent their active sites [54]. In addition, functionally linked proteins can interact with a common interacting partner via similar peptide patterns [54]. All of these premises strongly suggest that function-specific protein motifs can also be exploited for predicting protein functions. Development of a tool for discovering function-specific protein motifs with a near-zero false–positive rate,

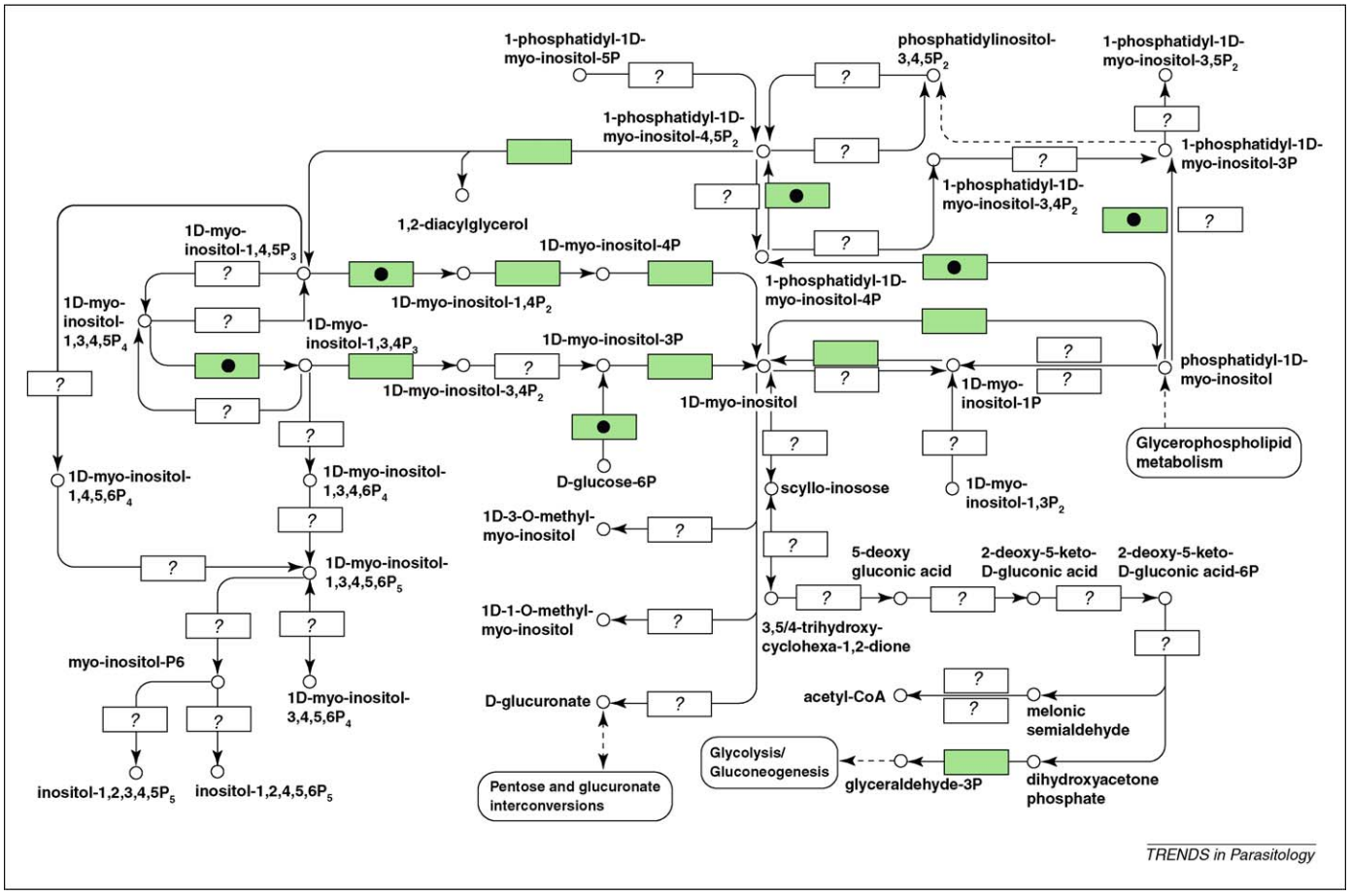

Figure 2. Inositol phosphate metabolism pathway and its known components in *Trypanosoma brucei*. Each box represents one of the enzymes of the consensus inositol phosphate metabolism pathway, as determined by KEGG. Some genes are represented by more than one box as they encode enzymes that can catalyze several reactions. Light green boxes represent enzymes for which at least one homolog is known in *T. brucei*. Question marks indicate enzymes that lack an obvious homolog in *T. brucei*. For example, although the enzyme that converts 1D-myo-inositol-1P to 1D-myo-inositol is known, no enzyme for generation of 1D-myo-inositol-1P has been found in the genome of *T. brucei*. Light green boxes that are marked by black circles show conserved enzymes for which participation in inositol phosphate metabolism can also be predicted by the combination of codon usage and regulatory motifs (i.e. the overlap of KEGG annotations and our predictions). Three genes that are known to be involved in other pathways were also among the predicted inositol phosphate metabolism enzymes (false positives), suggesting an estimated precision of approximately 62%. This predictor was also able to find 31 hypothetical proteins that are probably involved in this pathway. Once their exact function is known, these proteins can potentially fill the holes (question marks) highlighted in this figure. This figure is drawn based on KEGG Pathway 'tbr00562'.

similar to what FIRE can do for nucleotide motifs, can be a great step for computational annotation of proteins.

Genome-wide expression profiling of genes has recently opened other ways for gene function prediction, yet based on experimentally derived expression patterns. For example, it has been shown recently that an in-depth analysis of mRNA levels in *T. brucei* during differentiation process can reveal function-specific variations among the expression patterns of genes [55]. Genes can then be clustered based on their expression patterns, often resulting in groups of biologically related genes. Each group can have a mixture of characterized and uncharacterized genes; the functions of the uncharacterized genes can thus be predicted based on the functions of the characterized genes within the same group. Combining the results of such genome-wide experiments with sequence-based computational approaches that are described here will secure a more accurate and more complete functional annotation of the genome.

Homology-based identification of physical interactions

Rosetta stones [16], interolog mapping [17] and phylogenetic profiling [20] are among the most prominent methods used for homology-dependent prediction of physical interactions. Interactions predicted using these methods are detected solely among conserved proteins; however, the results of these methods can be combined with the results of homology-independent annotation methods to also include trypanosomatid-specific proteins. These methods can be combined by several means such as naive Bayesian networks. This not only enables us to predict interactions among non-conserved genes but also reduces the number of false positives and enhances the sensitivity of prediction for conserved genes.

Concluding remarks and future directions

The availability of genome sequences of several trypanosomatid parasites has boosted the hope of finding novel drug targets by computational analysis of these genomes. However, genome annotation of trypanosomatids is far from complete. A significant increase in the genome-wide functional annotation of trypanosomatid proteins can lead to better understanding of the biology of trypanosomatids and to identification of novel targets for therapeutics against trypanosomatids. The robust methodology that is described here can be adapted for functional annotation and drug target prediction in other parasites.

Pipelining the tools reviewed here in a single completely automatic platform, in which the output of each module can act as the input for downstream modules, will vastly expand the power and ease-of-use of the proposed analyses, making them available to every researcher with access to even limited computational facilities. The main input of this pipeline will be the genome sequence of the parasite. It will be able to generate 'gold standard' training sets automatically from the submitted genome sequence (i.e. based on known 'interactomes' in other well-studied organisms) for *in situ* training of each of its different computational modules. Alternatively, users can submit their own training sets at desired steps (i.e. based on experimental data). Using automatically generated gold standard training sets and user-defined training sets, a catalog of computationally predicted functional data can be created for all available parasite genomes, providing researchers with one of the most comprehensive databases specialized in parasite genomics.

Acknowledgements

The research in Salavati laboratory is supported by the Canadian Institutes of Health Research (CIHR) grant #94445 and the Natural Sciences and Engineering Research Council of Canada (NSERC) grant #328186. H.S.N. is supported by the Lloyd Carr-Harris Fellowship.

References

1 El-Sayed, N.M. *et al.* (2005) The genome sequence of *Trypanosoma cruzi*, etiologic agent of Chagas disease. *Science* 309, 409–415
2 El-Sayed, N.M. *et al.* (2005) Comparative genomics of trypanosomatid parasitic protozoa. *Science* 309, 404–409
3 Peacock, C.S. *et al.* (2007) Comparative genomic analysis of three *Leishmania* species that cause diverse human disease. *Nat. Genet.* 39, 839–847
4 Berriman, M. *et al.* (2005) The genome of the African trypanosome *Trypanosoma brucei*. *Science* 309, 416–422
5 Ivens, A.C. *et al.* (2005) The genome of the kinetoplastid parasite, *Leishmania major*. *Science* 309, 436–442
6 Sharan, R. *et al.* (2007) Network-based prediction of protein function. *Mol. Syst. Biol.* 3, 88
7 Slonim, N. *et al.* (2005) Information-based clustering. *Proc. Natl. Acad. Sci. U. S. A.* 102, 18297–18302
8 Beer, M.A. and Tavazoie, S. (2004) Predicting gene expression from sequence. *Cell* 117, 185–198
9 Eisen, M.B. *et al.* (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. U. S. A.* 95, 14863–14868
10 Zhou, X. *et al.* (2002) Transitive functional annotation by shortest-path analysis of gene expression data. *Proc. Natl. Acad. Sci. U. S. A.* 99, 12783–12788
11 Nariai, N. *et al.* (2007) Probabilistic protein function prediction from heterogeneous genome-wide data. *PLoS ONE* 2, e337
12 Shoemaker, B.A. and Panchenko, A.R. (2007) Deciphering protein–protein interactions. Part II. Computational methods to predict protein and domain interaction partners. *PLoS Comput. Biol.* 3, e43
13 Lu, L.J. *et al.* (2005) Assessing the limits of genomic data integration for predicting protein networks. *Genome Res.* 15, 945–953
14 Date, S.V. and Stoeckert, C.J., Jr (2006) Computational modeling of the *Plasmodium falciparum* interactome reveals protein function on a genome-wide scale. *Genome Res.* 16, 542–549
15 Jansen, R. *et al.* (2003) A Bayesian networks approach for predicting protein–protein interactions from genomic data. *Science* 302, 449–453
16 Marcotte, E.M. *et al.* (1999) Detecting protein function and protein–protein interactions from genome sequences. *Science* 285, 751–753
17 Yu, H. *et al.* (2004) Annotation transfer between genomes: protein–protein interologs and protein-DNA regulogs. *Genome Res.* 14, 1107–1118
18 Matthews, L.R. *et al.* (2001) Identification of potential interaction networks using sequence-based searches for conserved protein–protein interactions or "interologs". *Genome Res.* 11, 2120–2126
19 Enright, A.J. *et al.* (1999) Protein interaction maps for complete genomes based on gene fusion events. *Nature* 402, 86–90
20 Pellegrini, M. *et al.* (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc. Natl. Acad. Sci. U. S. A.* 96, 4285–4288
21 Date, S.V. and Marcotte, E.M. (2003) Discovery of uncharacterized cellular systems by genome-wide analysis of functional linkages. *Nat. Biotechnol.* 21, 1055–1062
22 Najafabadi, H.S. and Salavati, R. (2008) Sequence-based prediction of protein–protein interactions by means of codon usage. *Genome Biol.* 9, R87
23 Najafabadi, H.S. *et al.* (2009) Universal function-specificity of codon usage. *Nucleic Acids Res.* 37, 7014–7023
24 Benz, C. *et al.* (2005) Messenger RNA processing sites in *Trypanosoma brucei*. *Mol. Biochem. Parasitol.* 143, 125–134

25 Clayton, C. and Shapira, M. (2007) Post-transcriptional regulation of gene expression in trypanosomes and leishmanias. *Mol. Biochem. Parasitol.* 156, 93–101
26 Haile, S. and Papadopoulou, B. (2007) Developmental regulation of gene expression in trypanosomatid parasitic protozoa. *Curr. Opin. Microbiol.* 10, 569–577
27 Hotz, H.R. *et al.* (1997) Mechanisms of developmental regulation in *Trypanosoma brucei*: a polypyrimidine tract in the 3′-untranslated region of a surface protein mRNA affects RNA abundance and translation. *Nucleic Acids Res.* 25, 3017–3026
28 Irmer, H. and Clayton, C. (2001) Degradation of the unstable EP1 mRNA in *Trypanosoma brucei* involves initial destruction of the 3′-untranslated region. *Nucleic Acids Res.* 29, 4707–4715
29 Schurch, N. *et al.* (1997) Contributions of the procyclin 3′ untranslated region and coding region to the regulation of expression in bloodstream forms of *Trypanosoma brucei*. *Mol. Biochem. Parasitol.* 89, 109–121
30 Furger, A. *et al.* (1997) Elements in the 3′ untranslated region of procyclin mRNA regulate expression in insect forms of *Trypanosoma brucei* by modulating RNA stability and translation. *Mol. Cell. Biol.* 17, 4372–4380
31 Hehl, A. *et al.* (1994) A conserved stem-loop structure in the 3′ untranslated region of procyclin mRNAs regulates expression in *Trypanosoma brucei*. *Proc. Natl. Acad. Sci. U. S. A.* 91, 370–374
32 Vassella, E. *et al.* (2004) Expression of a major surface protein of *Trypanosoma brucei* insect forms is controlled by the activity of mitochondrial enzymes. *Mol. Biol. Cell* 15, 3986–3993
33 Vassella, E. *et al.* (2000) A major surface glycoprotein of *Trypanosoma brucei* is expressed transiently during development and can be regulated post-transcriptionally by glycerol or hypoxia. *Genes Dev.* 14, 615–626
34 Quijada, L. *et al.* (2002) Expression of the human RNA-binding protein HuR in *Trypanosoma brucei* increases the abundance of mRNAs containing AU-rich regulatory elements. *Nucleic Acids Res.* 30, 4414–4424
35 Webb, H. *et al.* (2005) Developmentally regulated instability of the GPI-PLC mRNA is dependent on a short-lived protein factor. *Nucleic Acids Res.* 33, 1503–1512
36 Mishra, K.K. *et al.* (2003) A negative regulatory element controls mRNA abundance of the *Leishmania mexicana* Paraflagellar rod gene PFR2. *Eukaryot. Cell* 2, 1009–1017
37 Purdy, J.E. *et al.* (2005) Regulation of genes encoding the major surface protease of *Leishmania chagasi* via mRNA stability. *Mol. Biochem. Parasitol.* 142, 88–97
38 Teixeira, S.M. *et al.* (1995) Post-transcriptional elements regulating expression of mRNAs from the amastin/tuzin gene cluster of *Trypanosoma cruzi*. *J. Biol. Chem.* 270, 22586–22594
39 Coughlin, B.C. *et al.* (2000) Amastin mRNA abundance in *Trypanosoma cruzi* is controlled by a 3′-untranslated region position-dependent cis-element and an untranslated region-binding protein. *J. Biol. Chem.* 275, 12051–12060
40 Mayho, M. *et al.* (2006) Post-transcriptional control of nuclear-encoded cytochrome oxidase subunits in *Trypanosoma brucei*: evidence for genome-wide conservation of life-cycle stage-specific regulatory elements. *Nucleic Acids Res.* 34, 5312–5324
41 D'Orso, I. and Frasch, A.C. (2002) TcUBP-1, an mRNA destabilizing factor from trypanosomes, homodimerizes and interacts with novel AU-rich element- and Poly(A)-binding proteins forming a ribonucleoprotein complex. *J. Biol. Chem.* 277, 50520–50528
42 Di Noia, J.M. *et al.* (2000) AU-rich elements in the 3′-untranslated region of a new mucin-type gene family of *Trypanosoma cruzi* confers mRNA instability and modulates translation efficiency. *J. Biol. Chem.* 275, 10218–10227
43 Bringaud, F. *et al.* (2007) Members of a large retroposon family are determinants of post-transcriptional gene expression in *Leishmania*. *PLoS Pathog.* 3, 1291–1307
44 McNicoll, F. *et al.* (2005) Distinct 3′-untranslated region elements regulate stage-specific mRNA accumulation and translation in *Leishmania*. *J. Biol. Chem.* 280, 35238–35246
45 Boucher, N. *et al.* (2002) A common mechanism of stage-regulated gene expression in *Leishmania* mediated by a conserved 3′-untranslated region element. *J. Biol. Chem.* 277, 19511–19520
46 Engstler, M. and Boshart, M. (2004) Cold shock and regulation of surface protein trafficking convey sensitization to inducers of stage differentiation in *Trypanosoma brucei*. *Genes Dev.* 18, 2798–2811
47 Colasante, C. *et al.* (2007) Regulated expression of glycosomal phosphoglycerate kinase in *Trypanosoma brucei*. *Mol. Biochem. Parasitol.* 151, 193–204
48 Zilka, A. *et al.* (2001) Developmental regulation of heat shock protein 83 in *Leishmania*. 3′ processing and mRNA stability control transcript abundance, and translation if directed by a determinant in the 3′-untranslated region. *J. Biol. Chem.* 276, 47922–47929
49 Quijada, L. *et al.* (2000) Identification of a putative regulatory element in the 3′-untranslated region that controls expression of HSP70 in *Leishmania infantum*. *Mol. Biochem. Parasitol.* 110, 79–91
50 Murray, A. *et al.* (2007) Regions in the 3′ untranslated region confer stage-specific expression to the *Leishmania mexicana* a600-4 gene. *Mol. Biochem. Parasitol.* 153, 125–132
51 Mao, Y. *et al.* (2009) Genome-wide computational identification of functional RNA elements in *Trypanosoma brucei*. *BMC Genomics* 10, 355
52 Elemento, O. *et al.* (2007) A universal framework for regulatory element discovery across all genomes and data types. *Mol. Cell* 28, 337–350
53 Chung, D.W. *et al.* (2009) Post-translational modifications in Plasmodium: more than you think! *Mol. Biochem. Parasitol.* 168, 123–134
54 Diella, F. *et al.* (2008) Understanding eukaryotic linear motifs and their role in cell signaling and regulation. *Front. Biosci.* 13, 6580–6603
55 Queiroz, R. *et al.* (2009) Transcriptome analysis of differentiating trypanosomes reveals the existence of multiple post-transcriptional regulons. *BMC Genomics* 10, 495