A systems approach towards functional annotation of the genome of *Trypanosoma brucei*

Hamed Shateri Najafabadi

Institute of Parasitology

McGill University, Montreal

August 2011

A thesis submitted to McGill University
in partial fulfillment of the requirements of the degree of
Doctor of Philosophy

Table of Contents

Abstract

The pathogenic species of trypanosomatids, including *Trypanosoma brucei*, *T. cruzi*, and *Leishmania* spp, cause serious human as well as animal diseases, with a very high incidence and mortality rate if untreated. Although the genome sequences of several trypanosomatids have been known for several years, many aspects of gene function and gene regulation are still unclear in these organisms. Most importantly, the lack of similarity between the majority of their genes and characterized genes of other organisms has limited our understanding of the gene functions in trypanosomatids. Not only the functions of many genes are unknown, the factors that are involved in their regulation are mostly uncharacterized. Trypanosomatids primarily rely on post-transcriptional programs for regulation of gene expression, and transcriptional regulation is of least importance. The genomes of these organisms harbour a large number of RNA-binding proteins with potential role in regulating mRNA stability and translation; however, the sequence specificity of these RNA-binding proteins and their function is mostly unknown. The focus of this thesis is on development of new methods for homology-independent functional characterization of genes in trypanosomatids, and deciphering the programs that are involved in their regulation. First, I describe a novel universal relationship between codon usage and gene function, and show the utility of this relationship for functional characterization of genes in various organisms, including trypanosomatids. This relationship most probably points to the role of codon usage in dynamic regulation of protein expression in different conditions, and helps the cell to adapt to new environments and conditions by synchronously regulating proteins with required functions. Then, I introduce a computational approach for identification of function-specific *cis*-acting regulatory elements, and demonstrate the utility of this approach for identification of potential regulatory elements in trypanosomatids, as well as for prediction of gene function based on the flanking regulatory sequences. I also show that combination of *cis*-regulatory elements and codon usage is a strong predictor of gene function in trypanosomatids. In addition to these methods, which can identify biological processes and pathways, a new method for identification of protein molecular functions based on short sequence signatures is introduced in this thesis. I show that this new method is able to identify function-specific protein short motifs that present functional sites on proteins, and demonstrate the utility of these motifs in predicting protein molecular function in trypanosomatids. In addition to these sequence-based approaches, I also explore the possibility of predicting trypanosomatid gene functions based on co-expression. I present the first co-expression network of *T. brucei*, which is constructed by combining several microarray datasets from different studies, and use it for predicting new components of several essential pathways in this organism. This analysis suggested the presence of a conserved post-transcriptional regulatory network in trypanosomatids, which encouraged us to develop a novel framework for identification of regulatory programs with high network-level conservation across multiple species. This framework revealed an extensive set of conserved regulatory programs in trypanosomatids, many of which could be validated using available expression datasets as well as our microarray profiles of chemical perturbations. The studies described here contribute significantly to functional annotation of genes in trypanosomatids, and identify the regulatory mechanisms that govern gene expression in these organisms. Furthermore, the introduced methods can be used for functional annotation of many uncharacterized genes and identification of gene regulatory programs in virtually all organisms with available genome sequences.

Résumé

Les espèces pathogènes de l'ordre des trypanosomatida, incluant *Trypanosoma brucei*, *T. cruzi*, et différentes espèces de *Leishmania* sont responsables de sérieuses maladies humaines et animales, avec une très forte incidence et taux de mortalité élevé lorsque non soignées. Bien que les génomes de plusieurs trypanosomatida soient disponibles depuis plusieurs années, de nombreux aspects de la fonction et de la régulation génique restent inexplorés chez ces organismes. De plus, l'absence de similarité entre la majorité de leurs gènes et les gènes caractérisés chez d'autres organismes a limité notre compréhension de la fonction de ces gènes chez les trypanosomatida. Non seulement la fonction de beaucoup de gènes est indéterminée, mais les facteurs impliqués dans leurs régulations ne sont, pour la plupart, pas encore caractérisés. Les trypanosomatida se reposent principalement sur des mécanismes post-transcriptionels pour la régulation de l'expression génique, et la régulation de la transcription n'a que peu d'importance. Les génomes de ces organismes hébergent un grand nombre de protéine se liant à l'ARN avec des rôles potentiels dans la régulation de la stabilité et de la traduction des ARNm. Néanmoins, les séquences spécifiques de ces protéines se liant à l'ARN et leurs fonctions restent principalement méconnues. L'objectif de cette thèse se situe au niveau du développement de nouvelles méthodes indépendantes de l'homologie pour permettre la caractérisation fonctionnelles de gènes chez les trypanosomatida, et de déchiffrer les mécanismes impliqués dans cette régulation. Premièrement, je décris une nouvelle relation universelle entre l'utilisation des codons et la fonction génique, et montre l'utilité de cette relation pour la caractérisation de gènes dans divers organismes, incluant les trypanosomatida. Cette relation pointe probablement vers un rôle de l'utilisation des codons dans la régulation dynamique de l'expression protéique sous diverses conditions, et aide la cellule à s'adapter à de nouveaux environnements et conditions en synchronisant la régulation des protéines avec les fonctions requises. J'ai introduis une approche computationnelle pour l'identification d'éléments cis-régulateurs fonction-spécifiques et démontré l'utilité de cette approche pour l'identification d'éléments régulateurs potentiels chez les trypanosomatida, ainsi que pour la prédiction de fonctions géniques basées sur les séquences régulatrices flanquantes. Je montre également que la combinaison d'éléments cis-régulateurs et l'utilisation des codons est un indicateur fort de la fonction génique chez les trypanosomatida. En plus de ces méthodes, qui peuvent identifier biologiquement des phénomènes et des voies métaboliques, une nouvelle procédure pour l'identification des fonctions moléculaires des protéines, basée sur de courtes signatures de séquences, est introduite dans cette thèse. Je démontre que cette nouvelle méthode est capable d'identifier de courts motifs protéiques fonction-spécifiques possédant des sites fonctionnels dont j'ai validé l'utilité pour identifier la fonction moléculaire de protéines chez les trypanosomatida. Outre cette approche basée sur les séquences, j'explore également la possibilité de prédire la fonction de certains gènes des trypanosomatida en me basant sur la co-expression. Je présente le premier réseau de co-expression de *T. brucei*, élaboré en combinant plusieurs jeux de données de *microarray* provenant de différentes études, et les utilise pour prédire de nouveaux éléments de multiples voies métaboliques essentielles dans cet organisme. Cette analyse suggère la présence de réseaux post-transcriptionels conservés chez les trypanosomatida, ce qui nous encourage à mettre au point un nouveau cadre expérimental pour l'identification de mécanismes régulateurs avec un fort niveau de conservation au sein de multiples espèces. Ce cadre expérimental a révélé une somme importante de mécanismes régulateurs conservés chez les trypanosomatida, dont beaucoup pourraient êtres validés en utilisant des données d'expression disponibles ainsi qu'avec des profils de perturbations chimiques de *microarrays*. Les études décrites ici contribuent significativement à l'annotation génique fonctionnelle chez les trypanosomatida, et permet d'identifier des mécanismes de régulation qui gouvernent l'expression génique de ces organismes. De plus, les méthodes introduites peuvent être utilisée pour l'annotation fonctionnelle de nombreux gènes non-caractérisés et l'identification de programmes de régulation génique dans virtuellement n'importe quel organisme dont le génome est disponible.

Contributions of Authors

Chapters 3, 4, 5, 7, and 8 of this thesis contain materials from previously published manuscripts (see references [1-5]). Reference [1] was co-authored by me and my supervisor, Reza Salavati. As the primary author, I contributed to method development, data preparation and analysis, and manuscript preparation. Reference [2] was co-authored by me, Hani Goodarzi, and Reza Salavati. As the primary author, I contributed to method development and experimental design, and performed the experiments on yeast, and also prepared and analyzed the data and wrote the manuscript. Hani Goodarzi contributed to experimental design and performed the experiments on *E. coli*. Reference [3] was co-authored by Yuan Mao, me, and Reza Salavati, in which Yuan Mao and I were the first co-authors. I contributed to method development, data preparation and analysis, and manuscript preparation. Yuan Mao contributed to data analysis for identification of conserved structural RNA elements. Reference [4] was co-authored by Reza Salavati and me. As one of the primary authors of this opinion paper, I contributed to analysis on combining codon usage and regulatory elements for pathway prediction, as well as to manuscript preparation. Reference [5] was co-authored by me and Reza Salavati. As the primary author, I contributed to method development, data preparation and analysis, and manuscript preparation. Reza Salavati provided intellectual input in all these studies, and contributed to experimental design, data analysis, and manuscript preparation.

Acknowledgements

I cannot overstate my gratitude to my supervisor, Reza Salavati. Over the past four years, his devotion has been an inspiration for me, and his knowledge has been reassuring. Reza’s trust in my judgment let me pursue the science that I cherished, and because of that, I enjoyed every moment of my research. His interdisciplinary vision of biology has created a lab where ideas can truly meet and mate, and I am grateful to him for accepting me as part of his lab.

I am also grateful to my advisory committee, Mathieu Blanchette and Robin Beech, who provided me with countless insights in our meetings and discussions. Also, I am thankful to past and present Salavati lab members for their support and friendship, and for the contribution they have had in the research I am presenting in this thesis. Especially, I thank Yuan Mao and Chad Macpherson for their extensive contribution to the research presented in Chapters 5 and 9, and Houtan Moshiri and Smriti Kala for their comforting support and help, without which the completion of this thesis was not possible. I also thank Rajashree Sen, Vaibhav Mehta, and Andrew Papadopoli for helping me in many occasions with my research.

I would also like to thank my great friends Hani Goodarzi and Kasra Hassani, whom I have known for more than ten years. Hani is an exceptionally talented colleague, and even though in the past four years he has been hundreds of miles away, we have collaborated on several projects, two of them presented in Chapters 4 and 6 of this thesis. Kasra is perhaps the greatest roommate one can imagine. His enthusiasm for science was what inspired me to pursue a career in research, and I will never forget all his support and encouragement when I started my PhD.

Also, I am grateful to Naghme Nikpour for being the most wonderful friend and the greatest inspiration. I cherish her wisdom, and appreciate her support in every hour of this endeavor over the past year. I thank all my friends in Montreal, especially Neda Moradin, Mitra Yousefi, Maryam Ehteshami, Golnaz Mousavi, and Nadjme Nikpour, and friends that, despite the distance, have never been far apart, especially Noori Torabi, Setareh Banihosseini, Reza Kalhor, and Manely Rashedan. And above all, I am deeply grateful to my family for their everlasting support, and for being beside me in every moment of the past four years, even though they were six thousand miles away.

To those who seek comfort in knowledge

1 Introduction

Trypanosoma brucei spp., *Trypanosoma cruzi* and *Leishmania* spp. are related parasitic protozoa with different life cycles and insect vectors (tsetse flies, reduviid bugs, and sandflies, respectively) and cause different diseases in human (and animals). *T. brucei* diverged most anciently, then *T. cruzi*, and most recently *Leishmania* within this trypanosomatid family [6]. The *T. brucei* group is an extracellular bloodstream pathogen that affects the central nervous system and causes human sleeping sickness (and nagana in cattle). *T. cruzi* is an intracellular pathogen that infects a wide variety of cells and causes Chagas disease that primarily manifests as severe cardiomyopathy. *Leishmania* spp. are intracellular pathogens in the mammalian host that target macrophages and cause a spectrum of diseases ranging from the lethal visceral form to the less severe cutaneous form of leishmaniasis. With a mortality rate of 50,000 individuals per year and annual loss of 2.1 million disability-adjusted life years (DALYs), leishmaniasis is the second most important parasitic infection after malaria, followed by sleeping sickness with 48,000 deaths per year (1.5 million DALYs per year)[7]. Chagas disease also causes 15,000 deaths and loss of 700,000 DALYs annually. The available drugs for these diseases are not ideal, since they are toxic, costly, and have invasive routes of administration[8-11]. Also, resistance to many of these drugs has already emerged; hence there is an urgent need for new drug development[12-15]. These diseases are also gaining increased attention in developed countries, particularly because of their transmission via blood transfusion and organ donation[16-20], as well as the high risk of infection among returning soldiers and immigrants[21-24].

The genome sequences of five trypanosomatids are published, including *T. brucei*, *T. cruzi*, *L. major*, *L. infantum* and *L. braziliensis*[25-29], and the genome sequences of four other trypanosomatids are available, including *L. mexicana*, *T. vivax*, *T. congolense*, and *L. donovani* (http://tritrypdb.org/tritrypdb/). Although these genome sequences, complemented by biochemical studies, have provided great opportunities to find new drug targets, their full potential is yet uncovered, given that the functions of the majority of genes encoded by these genomes are unknown. This scarcity of functional annotation primarily stems from lack of sequence similarity between many trypanosomatid genes and characterized genes of other organisms. Consequently, homology-based transfer of annotations from other genomes is simply impossible for more than half of trypanosomatid genes [30]. Not only the functions of many trypanosomatid genes are unknown, but also it is not clear how these genes are regulated in response to internal and external changes. Although it has been well established that the mRNA abundance is vastly regulated at the post-transcriptional level in trypanosomatids [31, 32], and that the transcriptome is remodeled extensively during the life cycle of these organisms [33-39], the regulatory factors that are involved in this process are mostly unknown.

The main objective of this research is to functionally characterize the trypanosomatid genes, and to understand how these genes are regulated developmentally or in response to external and internal stimuli, in order to obtain a systems-level understanding of the mechanisms that govern the biology of these parasites and help them adapt to their environment. Accordingly, two main specific aims are pursued in this research:

Functional annotation of coding sequences in trypanosomatid parasites – We have addressed this problem by developing several homology-independent computational methods for gene function prediction. These include sequence-based methods for prediction of protein-protein interactions and functional linkages, sequence-based methods for prediction of pathways and biological processes, novel methods for analysis of expression data in order to predict pathways, and sequence-based methods for prediction of protein molecular functions.

Characterization of non-coding functional elements and their role in gene regulation – We have developed new computational approaches for identification of non-coding functional elements in trypanosomatid genomes, and have used them in conjugation with experimental methods to identify *cis*- and *trans*-acting regulatory elements that govern mRNA stability and abundance in trypanosomatids and result in responsiveness against internal and external stimuli.

This research is particularly innovative in the approaches to achieve the scientific objectives, and the establishment of a functional genomics and bioinformatics pipeline that can be adapted to analysis of various pathways and biological processes in different organisms. Furthermore, the results contribute significantly to our knowledge of the biology of trypanosomatids, and provide new biological processes and genes that can be targeted for development of therapeutics against these parasites.

2 Literature Review

***The trypanosomatid parasites and their diseases* –** Despite the many similar biological aspects of the trypanosomatid species, they cause distinct human diseases whose vast prevalence, mortality rate and effect on normal life have rendered them major concerns in many developing countries. Two of the three subspecies of *Trypanosoma brucei*, *T. b. gambiense* and *T. b. rhodesiense*, cause human African trypanosomiasis (HAT); *Trypanosoma cruzi* is the causative agent of Chagas disease; and different *Leishmania* species are responsible for various forms of leishmaniasis [8]. With a mortality rate of 50,000 individuals per year and loss of 2.1 million disability-adjusted life years (DALYs), leishmaniasis is the second most important parasitic infection after malaria, followed by HAT with 48,000 deaths per year (1.5 million DALYs per year) [7]. Chagas disease also causes 15,000 deaths and loss of 700,000 DALYs annually. The current drugs for these diseases are unsuitable since they are very toxic and not very effective, or they have unpractical prices [8-11]. Also, resistance to these drugs has already emerged [12-15]. Hence, there is a need for new drug development.

***Human African trypanosomiasis* –** HAT is caused by *T. b. gambiense* and *T. b. rhodesiense*. The other subspecies of *T. brucei*, *T. b. brucei*, is not infectious to the human, hence is a safe model parasite for research. After infection, the parasite enters the bloodstream and transforms into the bloodstream trypomastigote, where it encounters the host immune response. However, the antigenic variation of the parasite, caused by either transcriptional switching between different expression sites (ES) or recombination of variable surface glycoproteins (VSG), causes immune evasion of the parasite [40]. For the same reason, development of a vaccine against HAT seems unlikely in the near future. The neurological symptoms of the disease emerge as the parasite passes the blood-brain barrier. The neurological damage caused by the parasite leads to coma and finally death, if untreated. The transmission of the parasite occurs via different species of the tsetse flies. The parasite transforms into trypomastigote procyclic form (PF) in the midgut of the fly, leaves the midgut and transforms to epimastigote, and reaches the saliva where it transforms to the infective metacyclic bloodstream form (BF) and can infect the mammalian host.

Chagas disease **–** Chagas disease is the result of infection with *T. curzi*. Following the infection, the metacyclic trypomastigotes invade cells of various tissues, including macrophages, smooth and striated muscle cells, and fibroblasts [41], where they transform to amastigotes that can multiply and transform again to infective trypomastigote form and infect other cells. The mechanism of entry is complex, and may involve lysosome recruitment activated by calcium-signaling pathway, which is specifically triggered by trypomastigote *T. cruzi*, or lysosome recruitment-independent mechanisms [42]. The trypomastigote can then escape from the formed parasitophorous vacuole and transform to amastigote. Shortly after infection, the acute stage of the disease begins which may be asymptomatic, followed by a drastic decrease in the parasite number. The chronic stage may start after a lag period of about 10-30 years (called indeterminate phase), after which sever symptoms arise such as cardiac damage, digestive damage and neurological disorders. If untreated, Chagas disease may be fatal.

Leishmaniasis **–** Ranging from the self-healing cutaneous infections to severe disfiguring mucocutaneous and lethal visceral disease [43], leishmaniasis can be caused by at least 21 species of *Leishmania* in human [44]. The fate of the disease is greatly influenced by the type of the immune response of the host to the infection (reviewed in [45]). After the infective metacyclic promastigotes are injected by the sandfly, the macrophages uptake them by phagocytosis, where they are delivered to phagolysosomes and differentiate into amastigotes which multiply and infect other cells [43].

Genomic features of trypanosomatids **–** The genome sequences of five trypanosomatids are published: *T. brucei*, *T. cruzi*, *L. major*, *L. infantum* and *L. braziliensis* [25-29]. Analysis of the genomes of the TriTryps, i.e., *T. brucei*, *T. cruzi* and *L. major*, has revealed that they have about 6100 genes in common, and about 1400, 3700 and 900 species-specific genes, respectively [26, 46]. Even at the first glance, the comparison of these genomes revealed interesting points. For example, several species-specific protein domains were identified, reflecting biological differences of these organisms, such as the presence of macrophage migration inhibitory factor only in *L. major* which may act in prevention of macrophage activation and subsequent destruction of the parasites, as *L. major* localizes into the phagolysosome of macrophages [26]. Comparison of the genomes of three *Leishmania* species also illustrated exciting differences, most notably the presence of an ortholog for the *T. brucei* Dicer-like protein TbDcl1 in *L. braziliensis*, implicating the possibility of the presence of RNAi machinery in *L. braziliensis* while absent from other *Leishmania* species [27]. However, the homology-based annotation of genes in the trypanosomatids is limited by their early divergence from other organisms with known genome sequences and hence, their poor similarity. For example, out of about 9,000 predicted and validated genes in *T. brucei*, about 5,000 do not have any reliable homolog in the sequenced genomes of non-trypanosomatid organisms[30]. The remaining 4,000 genes also cannot always be assigned a function since their homologs are also uncharacterized. Currently, only about 3400 *T. brucei* genes have any annotation other than hypothetical[30].

Genes in trypanosomatids are transcribed as polycistronic mRNAs that are further processed via trans-splicing, involving a polypyrimidine tract as the signal for spliced leader site [47]; this feature can be used for prediction of splice sites and, less confidently, polyadenylation sites from the genomic sequences, 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 [31, 32]. However, a few regulatory elements have been identified, all of which in the 3' UTR of developmentally regulated genes [48-71]. Some hints exist suggesting that elements in other regions rather than 3' UTRs may also play role in developmental regulation of expression [59], but none has been identified yet.

Computational annotation of the genome – In addition to the direct search for characterized homologs of a gene, other methods have been established by which gene functions can be inferred. Network-based approaches exploit the observation that proteins with similar functions usually interact with each other, therefore cluster together in the network of protein-protein interactions. Several different approaches have been exploited to assign functions based on protein-protein interactions, reviewed in [72]. Alternatively, genes can be clustered based on their expression patterns [73]; it is well established that genes with similar expression patterns have similar functions [74-76]. Also, functions can be assigned based on protein motifs. However, it is shown that combination of interaction network, expression patterns and protein motifs is superior to any of them alone, although interaction network alone contributes to about 85% of the predicted GO terms [77]. As the genome-wide interaction network is the most informative indicator of functional linkages between proteins, it is critical to obtain such a network. In the absence of experimental data, computational methods have been used to predict protein-protein interactions (reviewed in [78]). These methods can also be used for prediction of functional linkages between proteins instead of physical protein-protein interactions. Combination of these methods has proved powerful in computational modeling of interaction networks and functional linkages [79-81]. However, many of the prominent current methods rely on the presence of homologs in other species [82-87], limiting their application to conserved genes.

Current state of drug targets in trypanosomatids – Data provided by the genome sequences of TriTryps, complemented by biochemical studies, have provided opportunities for finding new drug targets. An ideal drug target should be present in all disease-causing trypanosomatids, either absent from the mammalian host or sufficiently different from its counterpart in the mammalian host, and essential for growth, survival or the pathogenicity of the parasite in the mammalian stages of its life cycle [8]. Other considerations are druggability, assay feasibility and resistance potential [88].

Being a monophyletic group of organisms [89], trypanosomatids share peculiarities distinguishing them from other well-known organisms, providing interesting targets for therapeutics. For example, (i) the sequences of most trypanosomatid mitochondrial pre-mRNAs are changed extensively through RNA-editing by addition and/or deletion of uridine nucleotides [90]; (ii) the mitochondrial DNA constitutes a major portion of the genome, and is usually structured into tens of maxicircles and hundreds of minicircles. This complex structure is necessarily accompanied by a set of complex processes for replication and segregation [91, 92]; (iii) mRNAs are transcribed as polycistronic clusters up to tens of kb long [93], all of which are processed by trans-splicing as an integral step of maturation [47]; and (iv) the major part of glycolysis is compartmentalized into specialized peroxisomes called gylcosomes [94].

In addition, several biochemical pathways have been proposed as potential drug targets in trypanosomatids. *Fatty acid biosynthesis* is an essential pathway; trypanosomatids use microsomal elongases to synthesize fatty acids *de novo*, whereas other organisms use elongases to make long-chain fatty acids even longer [95, 96]. It has been shown that inhibition of elongase pathway in the procyclic form of *T. brucei* results in growth defect [95]. Considering the importance of acyl and alkyl chains in various proteins and glycoconjugates of all trypanosomatids in different life stages, it will be probably not surprising to observe that elongases are also essential in *T. cruzi* and *Leishmania* spp [95]. Serving as another potential drug target, *Glycosylphosphatidylinositol (GPI) biosynthesis* is essential for the survival of *T. brucei* in the bloodstream stage [97, 98], perhaps due to the essentiality of the GPI-anchored VSG coat for cell morphology or due to the loss of the essential GPI-anchored trypanosome transferrin [99]. Although GPI biosynthesis pathway between *T. brucei* and mammals share common features, significant differences allow specific inhibition of this pathway in *T. brucei* [99], confirmed by analogues of a GPI intermediate that are toxic to *T. brucei* but not human cells [98]. *Ergosterol and isoprenoid biosynthesis pathways* have also been suggested as potential drug targets [8], especially that viability and proliferation of *T. cruzi* in all life stages requires specific sterols produced via these pathways [100-102]. Trypanosomatids also require salvage of folate and pteridines from their host as they are auxotrophic for them . Since rapidly dividing parasitic cells rely heavily on the availability of these compounds for *pyrimidine biosynthesis*, this pathway seems a promising drug target, as has been shown in the case of malaria parasites [104]. However, several factors, including the presence of pteridine reductases (PTR) in the trypanosomatids [105] that can bypass the dihydrofolate reductase (DHFR) inhibition, cause resistance to conventional antifolates, emerging the need for new solutions such as simultaneous targeting of both DHFR and PTR [106]. The *trypanothione redox metabolism* is another potential target. Trypanothione system is specific to trypanosomatids and replaces the nearly universal gluthatione system [107]. Enzymes of trypanothione redox pathway are suitable drug targets, especially the widely studied trypanothione reductase that seems to be essential in all trypanosomatids [108]. Targeting these enzymes not only can lead to parasite death *per se*, it can reduce resistance against other drugs. Despite the conservation of their active site residues between the host and the parasite [109], several enzymes of *carbohydrate metabolism* are also thought to be potential drug targets [110]. Glycolysis is specifically interesting as it is the only source of ATP at the bloodstream stage of the *T. brucei* life cycle. Galactose metabolism is another potential target, since *T. brucei* does not have a scavenger for the uptake of galactose from the host environment, while it is a crucial building block of VSG in the bloodstream form. Galactose-epimerase which converts glucose to galactose has been shown to be essential in both the bloodstream form and the procyclic form of *T. brucei*, rendering it a possible target candidate [111, 112]. Similarly, inhibition of mannose biosynthesis in *Leishmania mexicana* results in loss of virulence [113, 114], supposedly due to defective glycosylation of proteins involved in the establishment of infection in macrophages [115]. Other putative drug targets in trypanosomatids include *protein farnesyl transferases* [116-118], *cysteine proteases* [119], *N-myristoyltransferase* [120], *polyamine metabolism* [121], *purine salvage* [122, 123], *protein kinases* [124-126], *DNA topoisomerases* [127-129], and *RNA-editing enzymes* [130].

Although the essentiality of these pathways have been established in trypanosomatids, not all their constituting enzymes have been identified, mainly due to the lack of significant similarity with their known counterparts in other organisms. Identification of other enzymes involved in these pathways will reveal novel drug targets, especially since these ‘other’ enzymes will most probably be different from host enzymes and, hence, amenable to specific chemotherapeutic targeting.

3 Predicting physical and functional linkages among proteins based on codon usage

Knowing the physical and functional interactions among proteins, we can predict the functions of uncharacterized proteins based on the functions of their neighbors in the interaction network. This chapter explains a novel method for prediction of such networks based on codon usage of protein-coding genes, and appeared as an article in Genome Biology in 2008 [1]. In this chapter, we introduce a novel approach to predict interaction of two proteins solely by analyzing their coding sequences. We found that similarity in codon usage is a strong predictor of protein-protein interactions and, for high specificity values, is as sensitive as the most powerful current prediction methods. Furthermore, combining codon usage with other predictors results in 75% increase in sensitivity at a precision of 50%, compared to prediction without considering codon usage.

3.1 Background

The need for transforming the growing amount of biological information to knowledge has recruited several disciplines that, by means of experimental and computational approaches, aim to decipher functional linkages and interactions between proteins [72, 131]. Current computational methods for predicting protein-protein interactions demand data that, compared to the huge amount of available genomic sequences, are scarce. Only in a few organisms features such as essentiality, biological function and mRNA co-expression of genes have been partially determined. Also, a combination of different homology-based predictors, including phylogenetic profiles [86], Rosetta stone [82] and interolog mapping [83], provides incomplete information about interactions of only one-third of all *Saccharomyces cerevisiae* proteins. Hence, a method to identify protein-protein interactions solely on the basis of gene sequences would significantly expand the ability to predict interaction networks.

A few studies exist on prediction of protein-protein interactions based only on amino acid sequence information [132-134]. However, the highest specificity reported in these studies is 86%. Considering the number of possible protein pairs in a genome consisting of no more than 6000 protein-coding genes, this level of specificity means an unbearable number of 2.5×10^6 false positives. These studies consider the protein sequences, ignoring the plethora of information that exists in their coding sequences. The unsatisfied demand for reliable sequence-based prediction of protein-protein interactions encourages exploration of relevant sequence features in the genome instead of the proteome.

It has been widely acknowledged that codon usage is correlated with expression level [135]. In addition, it has been shown that codon usage is structured along the genome [136], with near neighbour genes having similar codon compositions. Some function-specific codon preferences have also been hypothesized based on selective charging of tRNA isoacceptors [137] and have been confirmed experimentally [138]. Based on these premises and considering that similarity in mRNA expression pattern and biological function, along with physical gene proximity, are powerful predictors of protein-protein interactions [79], codon usage can be considered as a potential candidate for analysis. The coevolution of codon usage of functionally linked genes has been explicitly reported before [139, 140]. These studies suggest that the codon adaptation index [141] of functionally related proteins change in a coordinated fashion over different unicellular organisms. However, identification of this coordination between two genes needs the presence of orthologues in several organisms; hence, many species-specific genes, which are usually the hot spots of attraction for biologists, are excluded. Also, there are genes with very low variation in the codon adaptation index over different organisms [139], for which this kind of analysis is unreliable.

In this paper, we will show that codon usages of functionally and/or physically linked proteins in an organism contain enough information to enable us to detect interacting protein pairs, even in the absence of homologues in other organisms. Furthermore, we will show that our method is several times more sensitive than tracking the coordinated changes of codon usage over different organisms, and in fact is one of the best methods for identification of protein-protein interactions.

3.2 Results and Discussion

Here we consider three different organisms: *S. cerevisiae*, *Escherichia coli* and *Plasmodium falciparum*. *S. cerevisiae* is a eukaryote with moderate coding G+C content (39.77%), while the genome of *P. falciparum* has an extremely low coding G+C content (23.8%), and *E. coli* is a prokaryote with moderate coding G+C (52.35%). For each organism, a positive and a negative gold standard set of protein pairs were defined, where a positive gold standard set comprises ORF pairs that, based on previous reports, encode proteins that interact with each other (either as members of the same protein-complex or as functionally linked proteins), and negative set consists of ORF pairs whose products do not interact with each other (Table 3-1). It should be noted that the highest resolution of our gold standard positive datasets is protein-complex. Given each ORF pair, we calculated the value of $d_{ij}(c)=|f_i(c)-f_j(c)|$ for each codon, where $f_i(c)$ and $f_j(c)$ are relative frequencies of codon c in ORF i and ORF j, respectively ($\Sigma_k f_i(c_k) = 1$ and $\Sigma_k f_j(c_k) = 1$; k = 1,2,..64 indicates all 64 codons). Therefore, d_{ij} demonstrates the distance of two ORFs in terms of usage of codon c. We found that for almost all codons, distribution of d differed between positive and negative gold standard sets (Supplementary Figure 3-1). Generally, distribution of *d* shifts to smaller values for ORFs within the gold standard positive set, indicating that interacting ORFs are more similar in codon usage profile than non-interacting ORFs. However, this shift is marginal for each codon individually, which means that single codons are weak predictors of protein-protein interactions.

Table 3-1. Gold standard sets used in this study – Each set comprises only ORFs that could be associated to their genomic sequences using the names that were provided in the original references. Self-interactions were considered neither in training nor in testing process. GSTD: Gold Standard Dataset; P: Positive; N: Negative; MIPS: Munich Information Center for Protein Sequences [142]; KEGG: Kyoto Encyclopedia of Genes and Genomes [143].

Organism	GSTD	Ref.	#ORFs	#ORF Pairs	Comments/Details
S. cerevisiae	P	[79, 81]	732	3,400	Derived from MIPS [142] complex catalog. We excluded ribosomal proteins to avoid bias towards extreme codon usage similarity of their genes.
	N	[79, 81]	2,760	1,442,691	Pairs of proteins that are not localized in the same cell compartment. We excluded ribosomal proteins.
P. falciparum	P	[80]	352	7,689	Protein pairs within the same KEGG [143] pathway.
	N	[80]	354	27,367	Protein pairs with KEGG information, excluding pairs in gold standard positive set.
E. coli	P	[144]	2,196	7,063	Pull-down assay using a His-tagged ORF library.
	N	-	3,703	4,437,833	We compiled a set of protein pairs that were not in gold standard positive set, given that at least one protein from each pair was co-purified with an associate protein by Arifuzzaman et al. [144].

We divided the distribution of *d* for each codon into 50 intervals, for each of which we calculated the likelihood ratio, i.e., the fraction of positive gold standards occurring in that interval divided by the fraction of negatives occurring in that interval. Since the mutual information of *d* for each pair of codons was negligible, we combined these likelihood ratios using a naïve Bayes approach (see Supplementary Figure 3-2 and Supplementary Figure 3-3 for graphical representations). Although not all features were independent from each other (with statistical tests suggesting 10 to 16 independent components; see Supplementary Figure 3-4), we found that a naïve Bayesian network is more effective than a Bayesian network in which each variable node has one other parent node, perhaps because the increase of the parameters in the latter case causes partial over-fitting of the network. Using a tenfold cross-validation method, we evaluated the performance of this naïve Bayesian network in predicting protein-protein interactions. To do so, we divided the gold standard set into ten random segments; each time we used nine segments as the training set and calculated the combined likelihood ratios for each ORF pair in the remaining segment. We designate the method "PIC" (Probabilistic-Interactome using Codon usage).

Figure 3-1A summarizes the performance of PIC in *S. cerevisiae*, *P. falciparum* and *E. coli*. For all three organisms, codon usage is a strong predictor of protein-protein interactions. As an extremely G+C poor parasite with a highly biased codon usage [145], the case of *P. falciparum* is of special interest, showing that codon usage is a powerful tool for prediction of interactomes within a wide range of G+C compositions. Figure 3-1B compares the performance of PIC in yeast with three widely used predictive methods: interolog mapping [83], phylogenetic profiles [86] and Rosetta stone [82, 85]. At low rates of false positives, PIC is the most sensitive method, up to seven times more sensitive than the next best method, interolog mapping. Also, for higher rates of false positives, PIC is still more sensitive than interolog mapping and Rosetta stone approach. Figure 3-1B also compares PIC with a previous report on identification of protein-protein interactions based on codon adaptation index coevolution [139], illustrating up to eight times higher sensitivity for PIC (see Methods for the details of the analysis). Finally, for the sake of comparison, the predictive power of the absolute difference of CAI (codon adaptation index; see reference [141] for the definition of CAI and to compare it with PIC) between two genes is investigated, showing a very poor performance (Figure 3-1B).

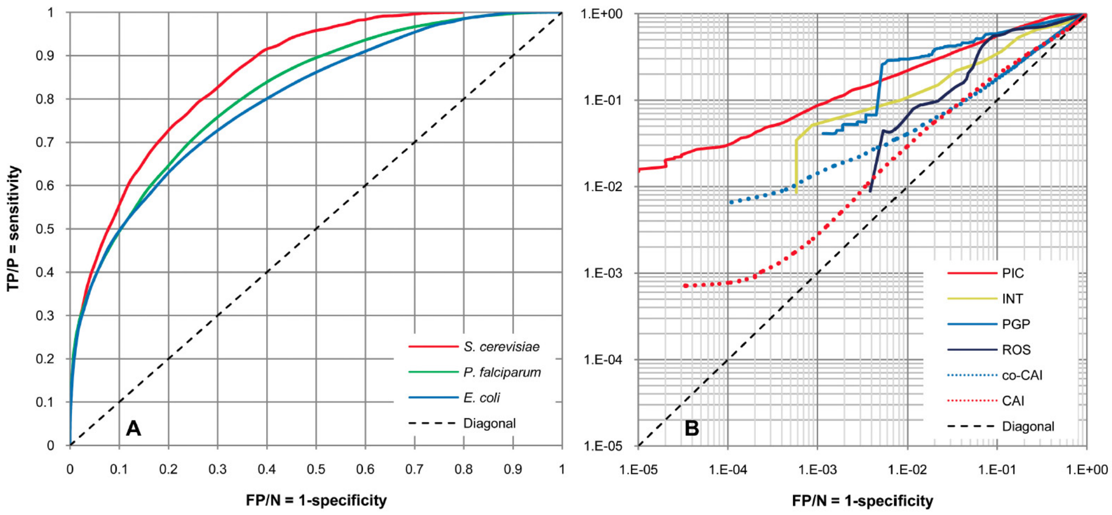


Figure 3-1. Results of protein-protein interaction prediction by PIC – (A) ROC (Receiver Operating Characteristic) curves of PIC for *S. cerevisiae* (red), *P. falciparum* (green) and *E. coli* (blue). **(B)** Comparison of ROC curves in yeast for PIC (red), Interolog mapping (INT, green), Phylogenetic profiles (PGP, blue), Rosetta stone (ROS, dark blue), CAI coevolution(co-CAI, blue dotted line) and absolute CAI value (CAI, red dotted line). The dashed line shows the diagonal. The same comparison is shown using the precision-recall curves in **Supplementary Figure 3-10**. For interolog mapping, phylogenetic profiles and Rosetta stone, data were retrieved from [146]. TP: true positive; P: positive; FP: false positive; N: negative. Positive and negative test sets are as indicated in **Table 3-1**.

It should be noted that the gold standard negative set that we used for *S. cerevisiae* is made of protein pairs that do not co-localize. Therefore, it may be possible that PIC recognizes subcellular localization of proteins instead of protein-protein interactions. To examine this, we compiled a set of protein pairs that localize within the same subcellular compartment. Then, we assessed the enrichment of interacting protein pairs and co-localized protein pairs in the positive predictions of PIC at different thresholds. As Figure 3-2 shows, the PIC predictions are rapidly enriched by true interacting proteins rather than proteins that are localized in the same subcellular compartment. We also compiled an alternative standard negative set by using pairs of proteins that have KEGG information [143], but do not share any KEGG pathway. Although this negative set is not as reliable as the main gold standard negative set that we used for the training and testing of PIC, it allows pairs of proteins that reside within the same subcellular compartment. The performance of PIC over this negative set was essentially the same as over the main gold standard negative set. For the other two studied organisms, *E. coli* and *P. falciparum*, the gold standard negative sets already contained co-localizing protein pairs.

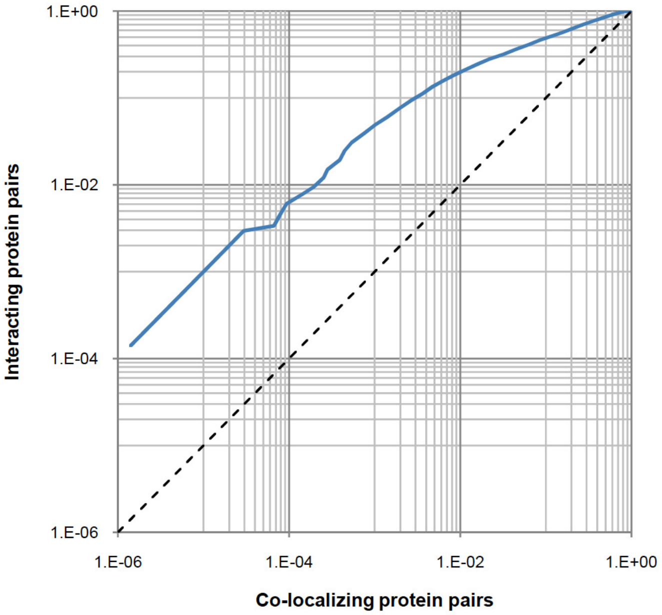


Figure 3-2. Enrichment of PIC predictions by interacting protein pairs versus protein pairs that co-localize – The horizontal axis shows the fraction of co-localizing protein pairs that match PIC predictions, and the vertical axis shows the fraction of the gold standard interacting protein pairs that match PIC predictions. Rapid enrichment of PIC with interacting protein pairs indicates that it detects protein-protein interactions rather than localization.

Although PIC considers the relative frequencies of codons in ORF pairs, it reflects not only synonymous codon usage, but also amino acid frequencies and ORF lengths. ORF length is reflected in PIC since stop codons are not omitted, and each ORF has only one stop codon. Therefore, the relative frequency of a stop codon in long ORFs is smaller than in short ORFs. We created three other probabilistic interaction networks of *S. cerevisiae* using RSCU [147], relative frequencies of amino acids, and ORF length, in order to examine the effect of each factor. We named these probabilistic networks PI-RSCU, PI-A and PI-L, respectively. RSCU is a measure of synonymous codon usage that is independent of amino acid composition (see reference [147] for the definition of RSCU and to compare it with the relative frequency of codon). RSCU as well as many other measures of synonymous codon usage are dependent on gene length, and result in biased values when the corresponding coding sequences are short [148]. In the worst case, when an amino acid is absent from a gene, it is impossible to calculate the RSCU for its corresponding codons. In the latter case, we treated the RSCU values of these codons as missing data, which can be easily handled by naïve Bayesian networks. In comparable sensitivities, the descending order of accuracy was PIC > PI-RSCU > PI-A > PI-L (see Supplementary Figure 3-5). This suggests a synergistic effect of each of these factors on the strength of PIC, with synonymous codon usage being the most important one. It should be mentioned that the length of the protein (PI-L) has a very marginal ability of distinguishing interacting from non-interacting pairs, and even this observed marginal prediction may be due to the bias of the gold standard positive set towards a certain range of protein lengths, as the length of a protein affects many experimental procedures such as successful cloning, etc.

PIC can easily be combined with other probabilistic approaches, such as PIP and PIT [81] (see the Methods section for combining two probabilistic interactomes). PIP is a probabilistic predicted network of *S. cerevisiae*, in which four datasets of genomic features are integrated: two datasets of biological functions, a dataset of mRNA expression correlation and a dataset of essentiality [81]. Jansen *et al.* [81] showed that, at comparable levels of sensitivity, PIP is even more accurate than PIE, a probabilistic network constructed by integration of four experimental datasets of the yeast interactome. They also combined PIP and PIE into PIT as one of the most comprehensive probabilistic networks of known and putative protein complexes in yeast. We integrated the results of yeast PIC and PIP, to see how their combination improves our power in *de novo* prediction of interactions.

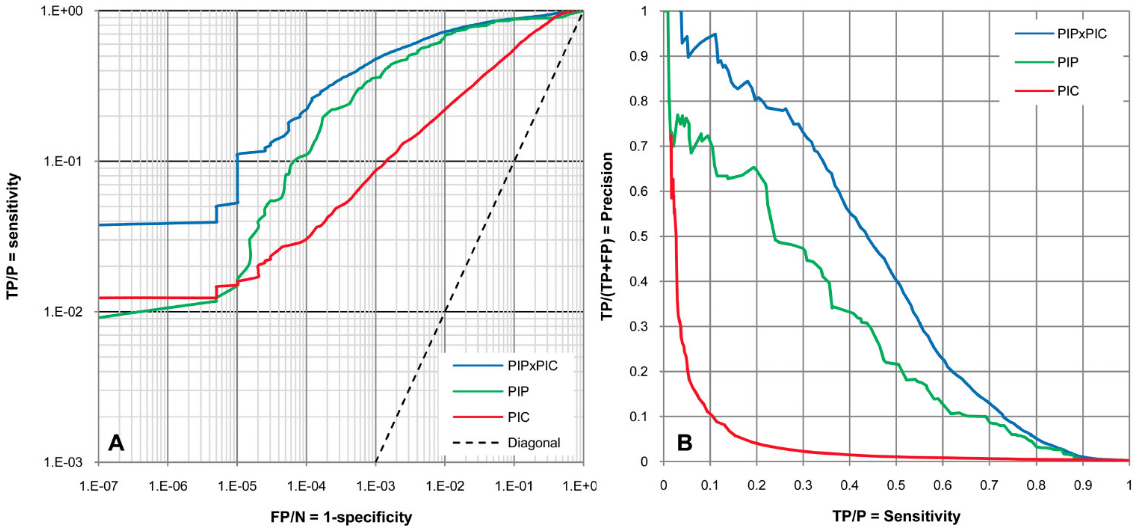


Figure 3-3. Comparison of performance in yeast for PIC, PIP and their combination – PIC is shown in red, PIP [81] in green and the combination of PIP and PIC (PIPxPIC) in blue. **(A)** ROC curves. Both axes are on log-scale. The dashed line shows the diagonal. **(B)** Precision-recall curves. TP: true positive; P: positive; FP: false positive; N: negative. Positive and negative test sets are as indicated in **Table 3-1**.

PIC, PIP [81] and their combination are compared in Figure 3-3. For false positive rates $<10^{-5}$, PIC is as sensitive as PIP, although in general PIP is far superior to PIC. More strikingly, combining PIP and PIC results in a four-fold increase in sensitivity when false positive rate is $<10^{-5}$ (after adding ribosomal proteins to the test set, a six-fold increase was observed). The combination of PIP and PIC remains the superior predictor for all false positive rates, and gets to a sensitivity of about 1.75 times that of PIP at a precision of 50%. Jansen *et al.* [81] used a likelihood threshold of 600 to cut an interaction network of *S. cerevisiae* out of PIP, referred here as PIP-Lcut$_{600}$. For comparable specificity, the combination of PIP and PIC is 1.5 times more sensitive than PIP-Lcut$_{600}$ (considering ribosomal proteins in the test set, the combination of PIP and PIC is 1.6 times more sensitive than PIP-Lcut$_{600}$; see Supplementary Figure 3-6). We also calculated the per-complex sensitivity of predictions for either PIP or combination of PIP and PIC, and observed that the combination of PIP and PIC outperforms PIP in every single complex as well (Supplementary Figure 3-7). Furthermore, we found that, compared to PIP, PIC in yeast is less biased towards certain biological functions (Supplementary Figure 3-8) as well as highly expressed genes (Supplementary Figure 3-9). However, it is evident that at least in the case of *P. falciparum*, PIC top-scoring interactions mainly belong to the ribosomal proteins. This reflects the very similar codon usage profiles of ribosomal proteins, most likely optimized for their efficient translation.

Finally, we combined PIT [81] and PIC to generate "PICT" which we propose as one of the most reliable probabilistic interactomes of *S. cerevisiae* (see Supplementary Figure 3-11 for precision-recall curves of PIT and PICT. PICT, accompanied by PIC for the whole genome of *S. cerevisiae*, is available online [149]). At a likelihood cutoff of 2×10^3, PICT has the same specificity as PIT-Lcut$_{600}$, while, after excluding promiscuous nodes (i.e., nodes each of which has $\geq$100 edges), includes 1306 more ORFs compared to PIT. Analysis of PICT-Lcut$_{2000}$ reveals many interesting interactions not present in PIT-Lcut$_{600}$. Some examples are represented below. We specifically consider complexes that were also examined by Jansen *et al.* [81], in order to provide a more detailed comparison between PIT and PICT. Note that the following interactions should be considered as complex co-memberships rather than direct physical interactions, since all the components of PICT are trained on protein complexes and not binary physical interactions of proteins. However, a direct physical interaction is also possible based on the closeness of proteins within the same complex.

While mammalian Pob3, an interacting partner of the nucleosome, has a high mobility group (HMG) for interaction with histones, yeast Pob3 lacks this domain [81]. Instead, in yeast, the HMG protein Nhp6 interacts with the nucleosome. PIT-Lcut$_{600}$ suggests that Nhp6A, an isoform of Nhp6, interacts with all nucleosome histones H2A, H2B, H3 and H4, which is highly unlikely considering the structure of the nucleosome. In addition, it has been shown that Nhp6 does not influence nucleosome reassembly; thus, it is unlikely for Nhp6 to interact with the H2A-H2B dimer [81]. In contrast to PIT-Lcut$_{600}$, PICT-Lcut$_{2000}$ only suggests an interaction between Nhp6A and HHT1 (H3), which is more congruent with the current models of nucleosome structure and assembly. PICT-Lcut$_{2000}$ also predicts a novel interaction between Nhp2, another HMG related protein, and H3 (Figure 3-4). Recently, affinity capture of Nhp2 has been shown to result in co-purification of histone proteins [150], corroborating the interaction of this protein with the nucleosome. PICT-Lcut$_{2000}$ also predicts the interaction of an uncharacterized ORF, YDL085C-A, with the nucleosome as well as with Nhp6A, which is consistent with previous reports showing the presence of GFP-fused YDL085C-A in the nucleus [151]. This example shows the potential of PICT, and codon usage in particular, to predict interactions of uncharacterized proteins, which should provide new insights into their probable functions.

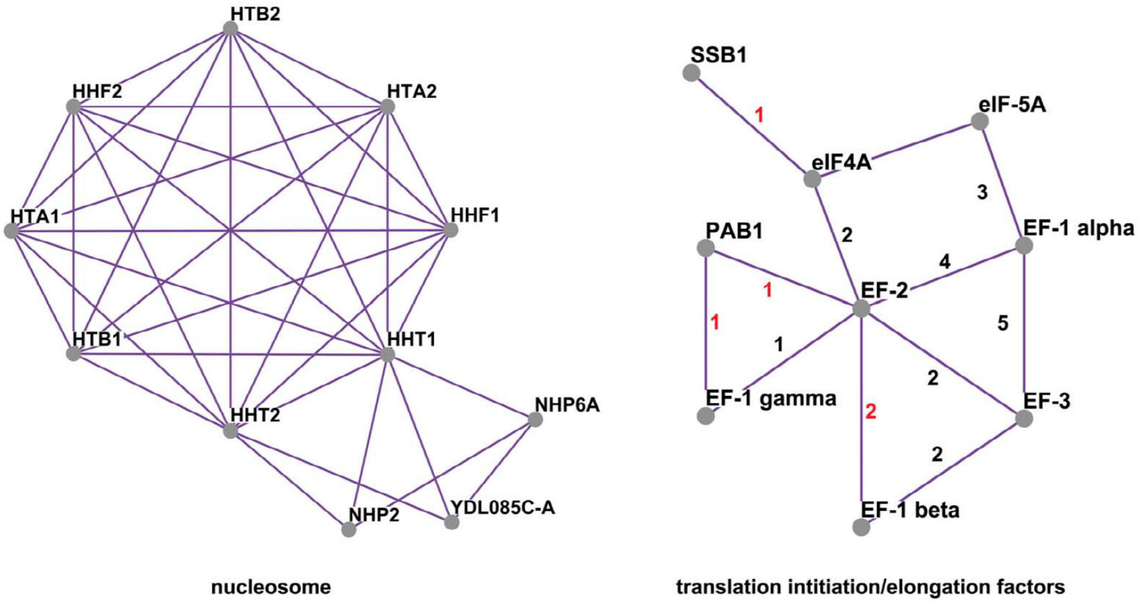


Figure 3-4. Two examples of complexes suggested by PICT-Lcut$_{2000}$ – In the case of translation initiation/elongation factors, only novel interactions (interactions absent from PIT-Lcut$_{600}$ [81]) are represented. A black number between two nodes stands for the reference in which the direct interaction of the two connected nodes is reported. A red number refers to the reference in which interaction of the two connected nodes with a third common protein is reported. 1: Gavin et al. 2006 [152], 2: Collins et al. 2007 [153], 3: Jao and Chen 2006 [154], 4: Jansen et al. 2003 [81], 5: Anand et al. 2003 [155].

Another example is the case of translation initiation/elongation factors. PIT-Lcut$_{600}$ fails to predict any interactions involving elongation factor 2 (EF-2). It also predicts only two interactions for EF-1α, with EF-1β and EF-1γ. Although PIT-Lcut$_{300}$ suggests a few more interactions for these proteins, higher rate of false positives in PIT-Lcut$_{300}$ renders them unreliable. PICT-Lcut$_{2000}$ predicts several interactions involving different elongation factors as well as initiation factors 4A and 5A, many of which have been recently confirmed by tandem-affinity purification experiments [81, 152-155]. Figure 3-4 shows a sub-graph of PICT-Lcut$_{2000}$ representing interactions among translation initiation/elongation factors that are not present in PIT-Lcut$_{600}$. A recent study [152] has shown that poly(A)-binding protein Pab1 interacts with EF-1α. Based on PICT-Lcut$_{2000}$, we anticipate that Pab1 interacts with EF-2 and EF-1γ as well. Also, we found an interesting interaction between the ribosome-associated molecular chaperone Ssb1 and eIF4A. Interaction of Ssb1 and eIF4G has already been shown by tandem-affinity purification [152]. Based on the close interaction of eIF4A and eIF4G, interaction of Ssb1 and eIF4A is reasonable.

RNase P complex represents another interesting example of PICT predictions. PICT-Lcut$_{2000}$ predicts six new interactions between RNase P complex and other proteins in yeast, neither of which exists in PIT-Lcut$_{600}$ or has been reported previously. Four interactions are with uncharacterized ORFs YKL096C-B, YDL159W-A, YKL183C-A and Q0255. Q0255 is likely to encode a maturase-like protein. It has been hypothesized that mitochondrial maturases participate in splicing by stabilizing some secondary or tertiary structure needed for splicing [156]. Their exact function, however, remains uncharacterized [157]. An interaction between RNase P complex and Q0255 implies the plausibility that this protein could contribute to maturation of ribosomal RNA and tRNA in mitochondria. According to PICT-Lcut$_{2000}$, HUB1 (Histone Mono-Ubiquitination 1) is another interacting partner of RNase P complex. Previous data have shown that HUB1 is a functional homolog of the human and yeast BRE1 proteins, and suggest that it mediates gene activation and cell cycle regulation through chromatin modifications [158]. In addition, chromatin remodeling in *Arabidopsis thaliana* seed dormancy has been proposed to be mediated by H2B monoubiquitination through HUB1 and HUB2 [158]. In agreement with this, the recently reported binding of human RNase P to chromatin of non-coding RNA genes and regulation of pol III transcription [159] could be mediated through HUB1-RNase P interaction. Another prediction of PICT-Lcut$_{2000,}$ the interaction of RNase P with CKB1, also corroborates this observation. CKB1 is a regulatory subunit of casein kinase 2, whose many substrates include transcription factors and all RNA polymerases. Again, this is consistent with the recently proposed role for RNase P in polymerase III transcription [159, 160].

We also noticed that PICT has the potential of providing new information about proteins that lack homology. For example, YAR068W is a fungal-specific gene, for which PIT has no interaction. This is while PICT predicts an interaction between this protein and a protein of the large subunit of mitochondrial ribosome.

3.3 Conclusions

PIC uses a naïve Bayesian network to combine the information provided by the frequencies of all codons, in order to predict protein-protein interactions. Given a set of independent features, naïve Bayesian networks can combine them in a way that minimizes the loss of information that usually occurs by the aggregation of several features. Depending on the training set that has been used, PIC can predict both complex membership (as in MIPS database or TAP-tagging experiments) and functional linkages between proteins (as in KEGG pathway database). Although we did not test the power of PIC for prediction of direct physical interactions between proteins, it is possible that it can be used for that purpose as well, since complex membership, functional linkage and direct physical interactions are all properties that are highly inter-correlated. We anticipate that integrating PIC with the current knowledge of protein interactions in different organisms will significantly increase the reliability and coverage of probabilistic interactomes. In the case of *Saccharomyces cerevisiae*, the results of PIC as well as its combination with PIT [81], referred to in this article as PICT, are provided online [161]. This study not only describes a novel method for *de novo* prediction of protein-protein interactions, but also suggests the plausibility of previously unseen evolutionary forces acting on codon compositions of genes within a genome. A few studies have taken into account the effect of protein-protein interactions on codon usage; however, these studies generally consider the unique features of codon composition of an ORF in regions that encode the interacting face of the protein compared to the rest of the ORF [162], not the direct relationship between codon usages of two interacting proteins. Characterization of evolutionary mechanisms shaping these relationships may lead to development of even more powerful methods for sequence-based prediction of interaction networks.

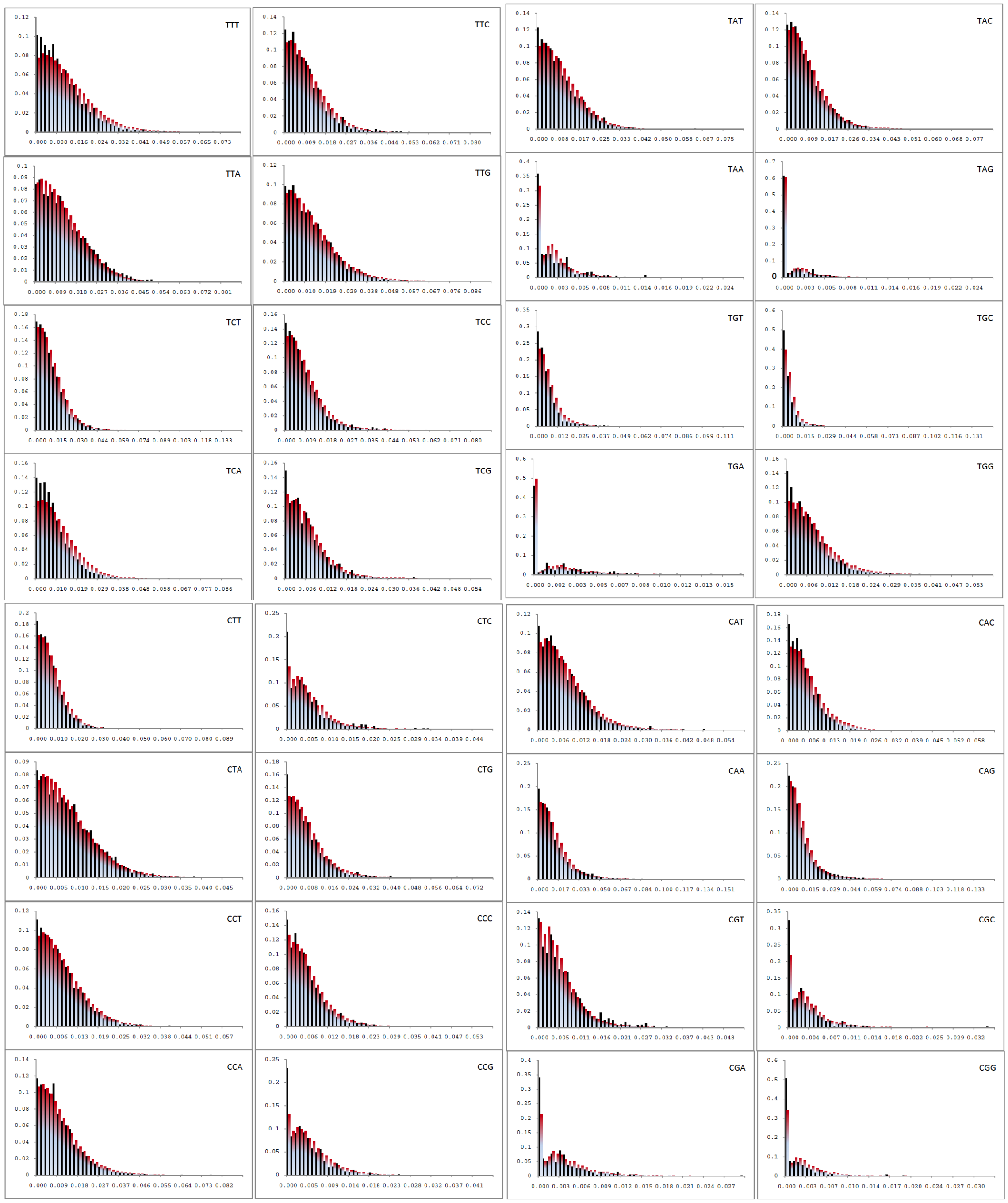
3.4 Methods

Genome Sequences – The genome sequences used were *S. cerevisiae* [163], *E. coli* [164] and *P. falciparum* [165].

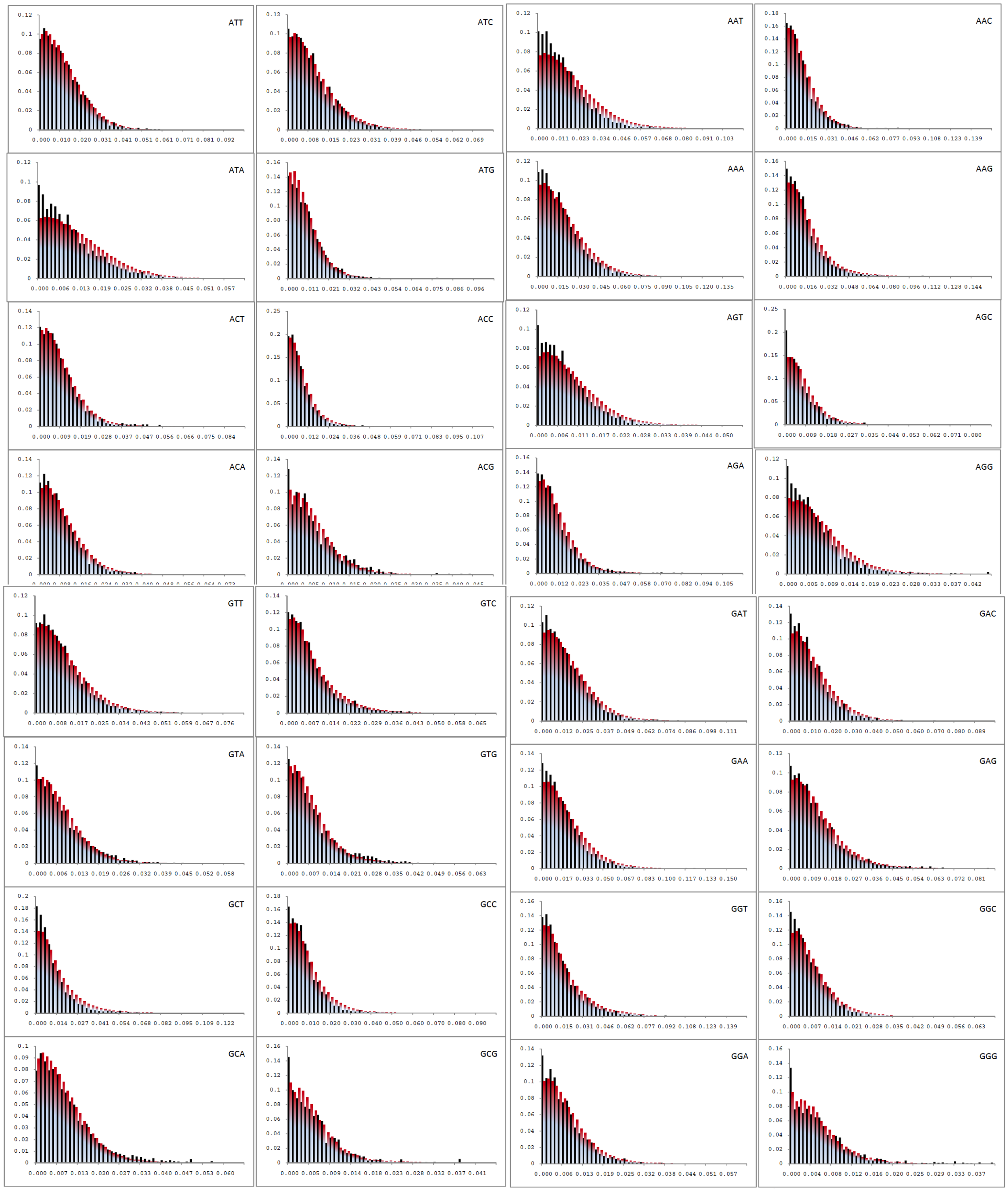
Analysis of genomic features – We used $d_{ij}(\zeta^k)=|\zeta^k_i-\zeta^k_j|$ to measure the distance of two genes *i* and *j* regarding feature ζ^k. In case of PIC, $\zeta^k=f(c_k)$, where $f(c_k)$ is the normalized frequency of usage of codon c_k, so that $\Sigma_k f(c_k) = 1$ ($1\leq k\leq 64$). For PI-RSCU, ζ^k=RSCU(c_k) (see [147]). For PI-A, $\zeta^k=f(a_k)$, where $f(a_k)$ is the normalized frequency of amino acid a_k ($1\leq k\leq 20$). For PI-L, ζ=L, where L represents the ORF length. To combine a set of features, naïve Bayesian network [79] is employed. Naïve Bayesian networks are most effective when they are used to combine independent features. We assessed independency of d_{ij} for two features *r* and *s* by means of mutual information [79], where $I[d_{ij}(\zeta^r); d_{ij}(\zeta^s)]<0.01$ was assumed not to influence the performance of the naïve Bayesian network. To combine two probabilistic networks, we multiplied the likelihoods each network assigned to each interaction.

Coevolution of CAI – We performed the same analysis as described by Fraser et al. [139], using the genome sequences of *S. cerevisiae*, *Saccharomyces paradoxus*, *Saccharomyces mikatae*, and *Saccharomyces bayanus* [166]. We used species-specific adaptation index to determine the CAI values by using the codon frequencies of the 20 most highly expressed genes. We assumed that the 20 most highly expressed genes in the four species are the same; hence, we used a previous report on mRNA expression in *S. cerevisiae* [167] to identify them. Addition of *Escherichia coli* in the analysis did not improve the results. We did not add more genomes because we would lose a portion of our gold standard sets, especially the negative gold standard set, due to the lack of homology for all genes among all genomes, resulting in non-comparable sensitivity/specificity values.

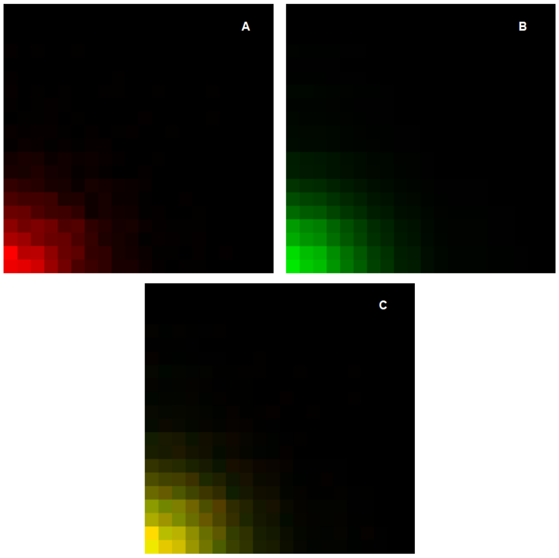
3.5 Supplementary Figures



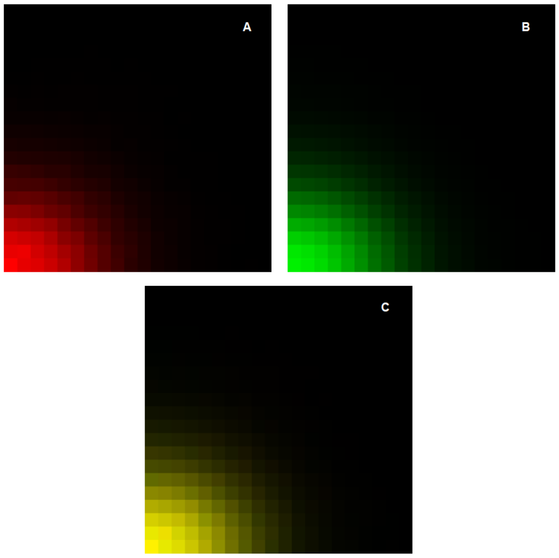
Supplementary Figure 3-1. Distribution of *d* for each codon in yeast – The value of *d* is demonstrated on the horizontal axis, whereas the vertical axis shows the density. Black bars represent distribution of *d* in positive gold standard set and red bars stand for the distribution of *d* in negative gold standard set. Figure continues on the next page.



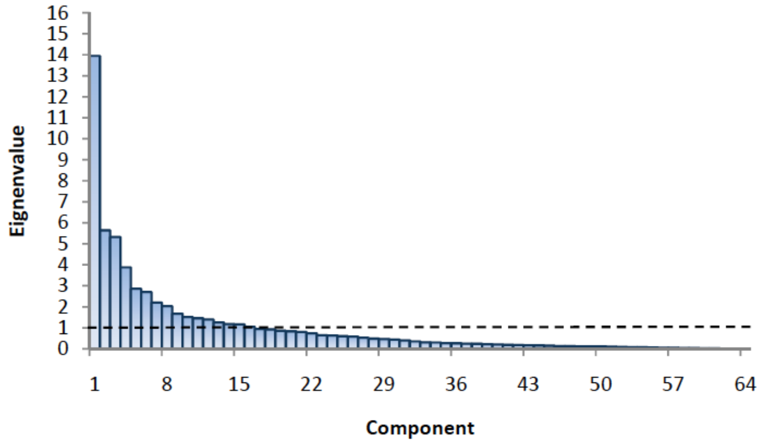
Supplementary Figure 3-1 continued



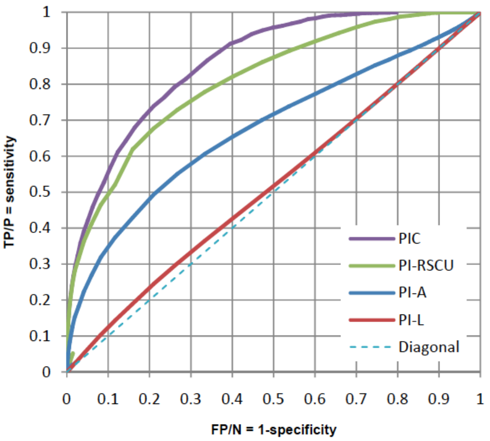
Supplementary Figure 3-2. Comparison of naïve Bayesian network and fully connected Bayesian network in yeast gold standard positive set – In each panel, the horizontal axis shows *d*(TTT) and the vertical axis shows *d*(TTC). Color intensity represents the probability of an interacting pair of proteins having the respective *d*(TTT) and *d*(TTC) values, predicted by either fully connected Bayesian network (shown in red, panel A) or naïve Bayesian network (shown in green, panel B). Panel C shows the combination of panels A and B (yellow).



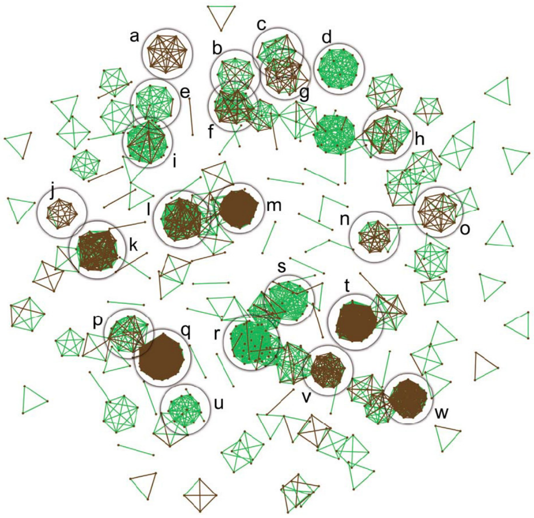
Supplementary Figure 3-3. Comparison of naïve Bayesian network and fully connected Bayesian network in yeast gold standard positive set – See the caption for **Supplementary Figure 3-2**.



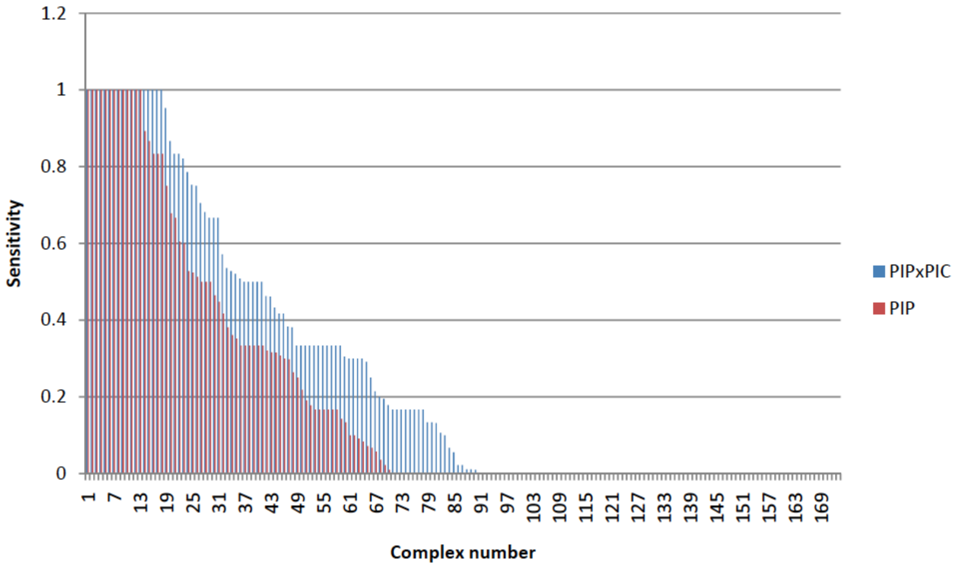
Supplementary Figure 3-4. Eigenvalues of different components resulted from principal component analysis of the interacting gene pairs in yeast – A 64-dimensional space was constructed consisting of all interacting pairs of genes from yeast gold standard set, so that dimension *k* represents $d(c_k)$ for each pair of genes. Then, this space was transformed using principal component analysis (PCA). In this figure, the components are in decreasing order of their eigenvalues, the first component being the first principal component of the transformed space. The first 16 components have eigenvalues greater than unity, suggesting the presence of 16 effective components based on the Guttman-Kaiser criterion (Guttman, L., 1954. Psychometrika XIX:149-61). However, the Scree test (Cattell, R.B., 1966. Multivariate Behavioural Research 1:245-76) suggests the presence of about 10 effective components. We also calculated an entropy-based number of effective codons for the collection of all coding sequences in *S. cerevisiae*, formulated as $N = e^{-\sum_{k=1}^{64} f(c_k)\ln\left[f(c_k)\right]}$, yielding in a value of 52.23. Both the effective number of components obtained from the Guttman-Kaiser criterion and the Scree test are small compared to the entropy-based number of effective codons, suggesting that the differences of codon frequencies in yeast contain redundant information.



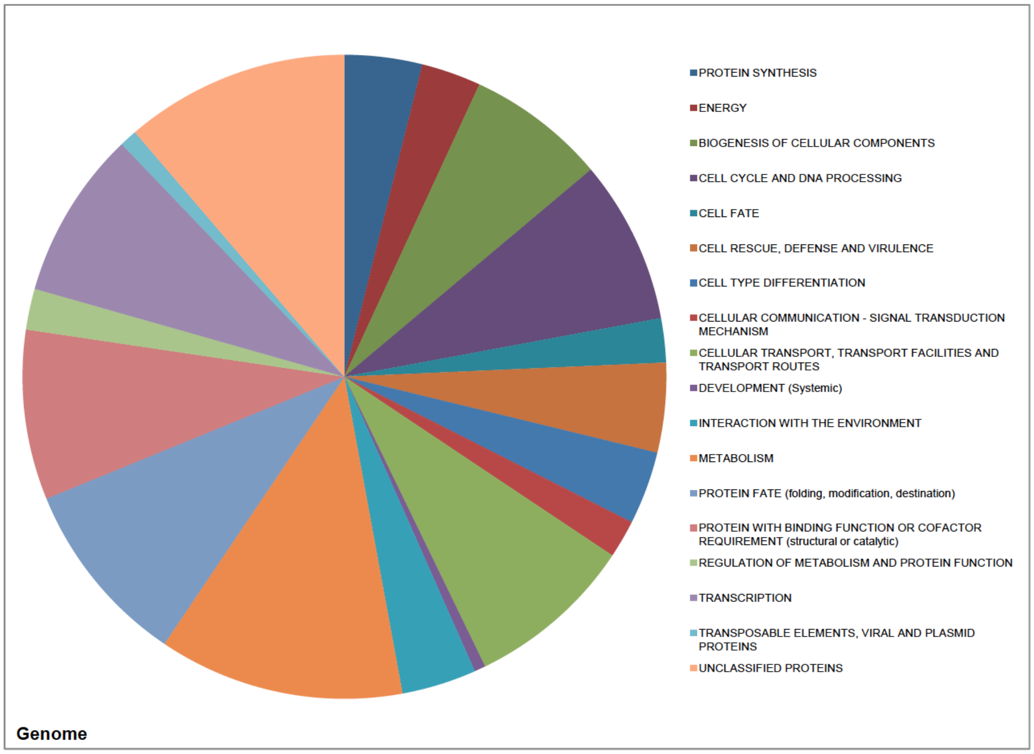
Supplementary Figure 3-5. Comparison of ROC curves for PIC, PI-RSCU, PI-A and PI-L



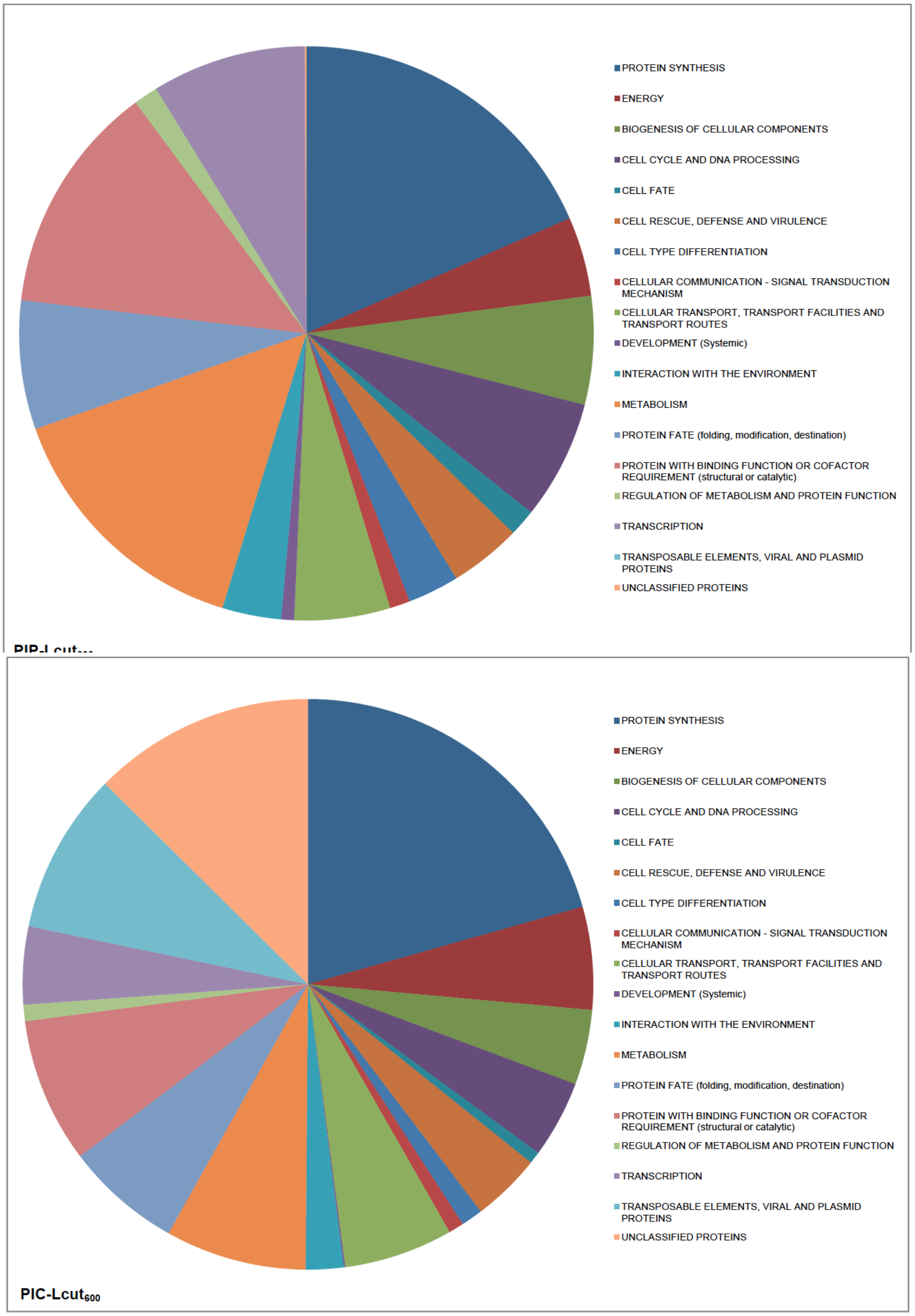
Supplementary Figure 3-6. Comparison of PIP×PIC and yeast gold standard positive set (including ribosomal proteins) – PIP×PIC (brown) covers 43% of the gold standard positives (green), while PIP has a coverage of 27% at the same specificity. (a) nucleosome; (b) COPI; (c) RNase MRP; (d) CCR4-NOT; (e) ER oligosaccharyl-transferase; (f) V-ATPase; (g) Cctring; (h) TFIIH; (i) small subunit of mitochondrial ribosome; (j) exocyst complex; (k) transcription; (l) pre-replication; (m) 40S ribosomal subunit; (n) ubiquinol cytochrome-c reductase; (o) cytochrome-c oxidase; (p) APC/C; (q) 60S ribosomal subunit; (r) RNA-polymerase II mediator; (s) SAGA complex; (t) proteasome; (u) TRAPP; (v) F1F0 ATP synthase; (w) large subunit of mitochondrial ribosome.



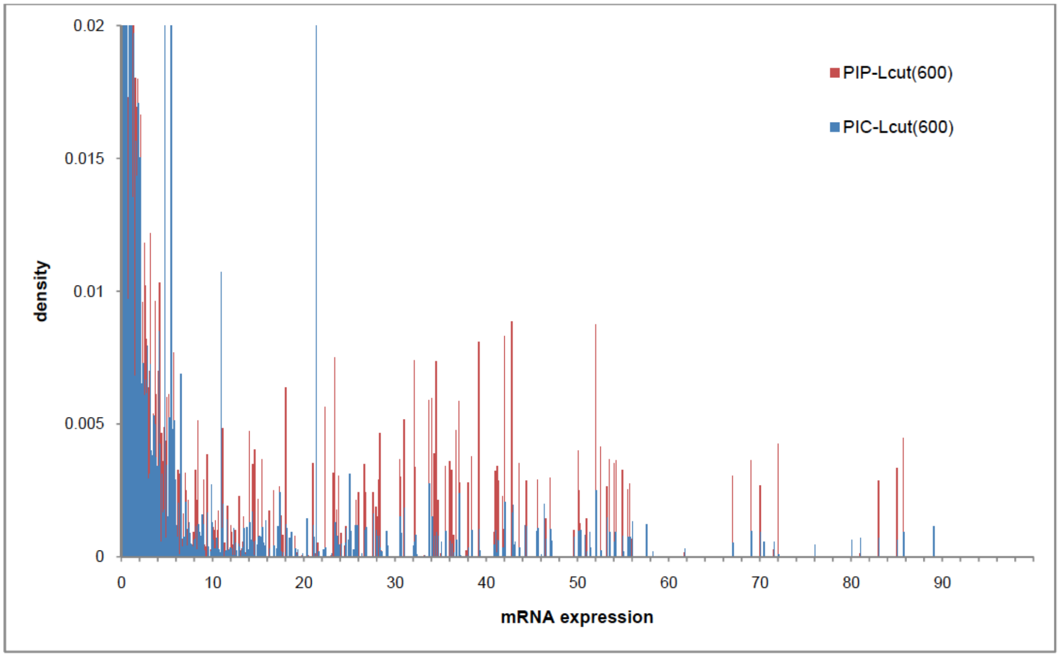
Supplementary Figure 3-7. Per-complex comparison of PIP and PIPxPIC – The sensitivity of each of these two methods is given for each complex. The complex numbers are in the descending order of the sensitivity of PIP.



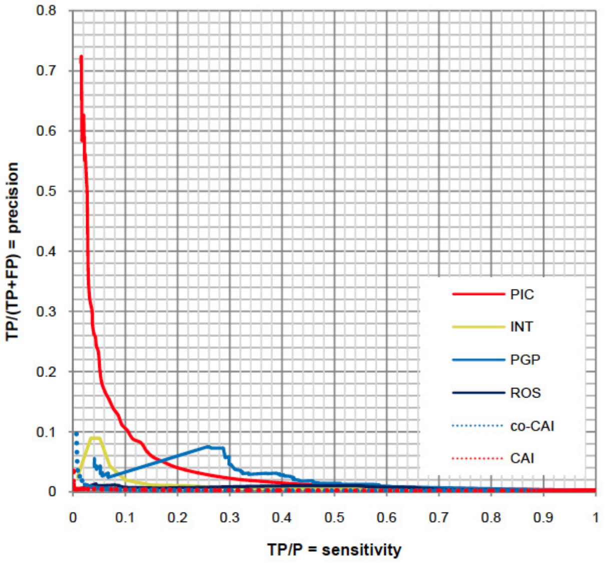
Supplementary Figure 3-8. MIPS functional category enrichment for yeast genome, PIP-Lcut600 and PIC-Lcut600 – Each slice reflects either the number of ORFs classified under the respective functional category (in the case of yeast genome), or the number of interactions involving at least one ORF from the respective category (in the case of PIP-Lcut600 and PIC-Lcut600). Both PIP-Lcut600 and PIC-Lcut600 are, compared to yeast genome, enriched by proteins involved in protein synthesis, even though the training set of PIC was deprived from ribosomal proteins. PIC-Lcut600 contains many interactions involving proteins of unclassified function, while PIP-Lcut600 includes few interactions of this kind, mainly because of its dependence on biological function as a predictor. Figure continues on the next page.



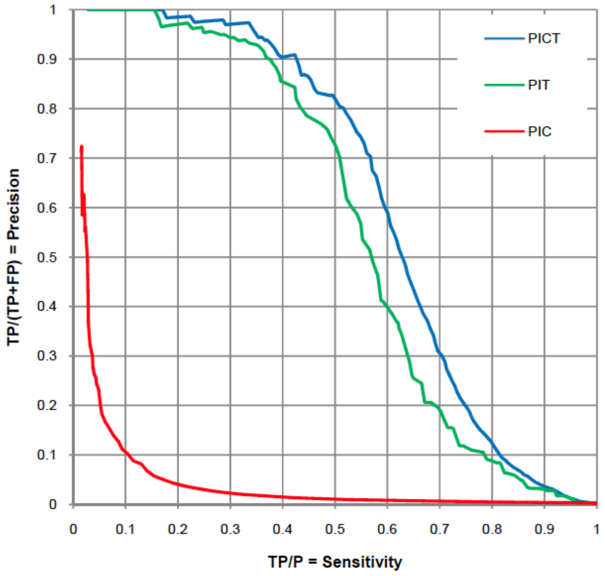
Supplementary Figure 3-8 continued.



Supplementary Figure 3-9. Distribution of mRNA expression levels in interactions predicted by PIP-Lcut600 (red) and PIC-Lcut600 (blue) for S. cerevisiae – PIC-Lcut600 shows less bias towards highly expressed ORFs compared to PIP-Lcut600. Expression data are retrieved from yeast reference mRNA expression set introduced by Greenbaum et al. (2002, Bioinformatics 18:585-96).



Supplementary Figure 3-10. Comparison of precision-recall curves in yeast for PIC (red), Interolog mapping (INT, green), Phylogenetic profiles (PGP, blue), Rosetta stone (ROS, dark blue), CAI coevolution(co-CAI, blue dotted line) and absolute CAI value (CAI, red dotted line).



Supplementary Figure 3-11. Comparison of precision-recall curves in yeast for PIC (red), PIT (green) and PICT (blue). At 50% precision, PICT is 12.5% more sensitive than PIT alone.

4 Universal function-specificity of codon usage

In the previous chapter, we showed that codon usage can predict physical and functional interactions among proteins in yeast, *Plasmodium falciparum*, and in *Escherichia coli*. Whether this observation holds in other organisms was, however, not addressed. Furthermore, the reason behind this relationship between synonymous codon usage and protein physical/functional linkages remained unclear. In this chapter, which appeared as an article in Nucleic Acids Research in 2009 [2], we propose that codon usage is ubiquitously selected to synchronize the translation efficiency with the dynamic alteration of protein expression in response to environmental and physiological changes. Our analysis reveals that codon usage is universally correlated with gene function, suggesting its potential contribution to synchronized regulation of genes with similar functions. We directly show that coexpressed genes have similar synonymous codon usages within the genomes of human, yeast, *Caenorhabditis elegans* and *E. coli*. We also demonstrate that perturbing the codon usage directly affects the level or even direction of changes in protein expression in response to environmental stimuli. Perturbing tRNA composition also has tangible phenotypic effects on the cell. By showing that codon usage is universally function-specific, our results expand, to almost all organisms, the notion that cells may dynamically alter their intracellular tRNA composition in order to adapt to their new environment or physiological role. Based on the notion of universal function-specificity of codon usage, in this chapter we also demonstrate the utility of codon usage in homology-independent prediction of the biological functions of genes based on their coding sequences, and we specifically show the application of this approach in *Trypanosoma brucei*.

4.1 Background

Genome-wide analysis of gene expression has been extensively used to study the mechanisms underlying the dynamic regulation of gene expression. Simultaneous changes in transcript levels across different conditions (i.e. environmental as well as spacio-temporal and genetic variables) have been widely used as a proxy for identifying sets of coregulated genes, thus revealing the common regulatory elements underlying these observed correlations [74, 168]. These regulatory elements can act at both transcriptional and post-transcriptional levels including mechanisms such as localization and stability of the transcript and enhancement or suppression of translation [32, 169]. Most studies have focused on the non-coding genome for finding such regulatory elements. While coding sequences have also been shown to contain elements that contribute to regulation of expression for a minority of genes [170-172], the relationship between the dynamic regulation of expression and the sequence of coding regions is not considered widespread among and within organisms.

A novel relationship between coding sequence and dynamic regulation of protein expression can be readily hypothesized from the observed variations in patterns of isoacceptor tRNA abundance and tRNA charging in different conditions and tissues [138, 173, 174]. For example, during amino acid starvation, unlike common tRNA species, the charged level of certain isoacceptor tRNAs cognate to rare codons remains high [137, 173]. Interestingly, these rare codons have been used in higher frequencies among genes involved in amino acid biosynthesis. Therefore, the high charged level of their cognate tRNA species can boost the amino acid biosynthesis pathway by supporting the high expression level of its enzymes [137]. Methylation of the wobble base of a tRNA in yeast has also been shown to affect its codon preference, enhancing levels of certain proteins [175].

Congruent with the observed tissue-specificity of tRNA composition [174], Plotkin *et al.* reported the presence of a tissue-specific codon usage in human genes [176], although Sémon *et al.* reject their hypothesis using a different statistical analysis and a richer database of tissue-specific genes [177]. However, many other studies indirectly suggest the act of selection on synonymous codons in human genome (reviewed in [178]); these models propose translational efficiency [179-181], mRNA stability [182-184], and splicing control [185] as mechanisms underlying such selection.

Despite all these studies, factors shaping codon usage of genes in many organisms, including human, are still not completely understood. For example, no significant evidence for the presence of selection on codon usage was found in 30% of the bacteria that were examined using a population genetics-based model [186]. A recent study has profoundly added to this complication by reporting that it is not the codon usage, but rather RNA structure that affects expression level of a protein [187].

More than thirty years ago, Garel [188] reported that the tRNA composition of silk gland in silkworm changes during the development of this organ in order to accommodate for the high rate of synthesis of fibroin which is rich in glycine, alanine, serine and tyrosine. In other words, silk gland cells try to synchronize the translation efficiency of fibroin with its required amount at each time by providing the tRNAs that carry these four amino acids. Matching tRNA composition with coding sequence may extend beyond amino acid usage of the proteins and include synonymous codon usage as well. Here, based on several rigorous statistical analyses of coding sequences from almost all available genomes, we suggest function-specificity of synonymous codon usage in a wide range of organisms. This implies that functional adaptation of tRNA content [188] may be a universal mechanism for synchronizing the translation efficiency with the dynamic, function-specific alteration of protein expression. In other words, rather than having a single set of optimal codons, organisms could harbor different sets that change depending on environmental conditions or physiological roles and are related to the functions that are most expressed at each of these conditions. We show that this hypothesis best explains the synonymous codon usage of genes across all domains of life. It also explains our recent observation that in three different organisms, *Saccharomyces cerevisiae*, *Escherichia coli* and *Plasmodium falciparum*, genes whose products interact either physically or functionally use similar codons [189]. Although not as comprehensive as our computational analysis, we also provide limited experimental data showing that differences in codon usage or variations in the tRNA content of the cell can result in varied responses to environmental changes, in terms of regulation of protein expression and cell phenotype.

4.2 Materials and Methods

Analysis of correlation between codon usage and function/expression pattern – Normalized frequency of each codon in each gene (f_c) was calculated as the usage of that codon divided by the usage of the amino acid it codes for. For each gene, we calculated the normalized frequencies of codons of an amino acid only if the usage of that amino acid was higher than a defined cutoff, in order to remove noisy measurements. The distance of normalized frequencies of each codon in each pair of genes was calculated, and the relationship between this distance value and likelihood of functional linkage/coexpression was assessed by Pearson correlation coefficient. Recently duplicated genes have high sequence identity and, thus, similar codon usages that may not necessarily reflect act of selection on gene sequence. In addition, these genes tend to cross-hybridize on expression arrays and appear as coexpressed genes. Therefore, to avoid biased analysis, paralogous genes were removed from each genome prior to calculating Pearson correlation coefficients. This was done by sequentially selecting two paralogous genes and randomly removing one of them, until the remaining genes contained no paralogs in the dataset [168]. Non-random distribution of f_c in each functional cluster or cluster of coregulated genes was assessed by mutual information of f_c across the genes of that cluster, with high mutual information values indicating non-random distribution and low mutual information values indicating random distribution.

Evaluating the effect of codon usage on the pattern of translation efficiency – The effect of codon usage on protein expression profile was assessed by measuring the amount of expression of two *lacZ* variants under 16 different conditions in yeast cells. One of these two *lacZ* variants was the genomic *lacZ* from *E. coli* K12-MG1665, and the other variant was a synthetic gene with the same protein sequence but extensively different codon usage. Both of these variants were inserted in pBridge (Clontech), and cloned in AH109 yeast strain (Clontech), thus having the same upstream and downstream sequences. The expression of each variant of *lacZ* in each growth condition was measured by a β-galactosidase assay using ortho-nitrophenyl-β-galactoside (ONPG) as substrate. The rate of conversion of ONPG to ortho-nitrophenol was measured by absorbance at

405nm. This rate, normalized for cell density (estimated by OD_{600}), was considered as the expression level of *lacZ*. Each of the experiments was performed in hexaplicate.

***Evaluating the ability of tRNA composition in conferring new phenotypes* –** A tRNA library was constructed with random combinations of 25 *E. coli* tRNA genes cloned into pBAD18 in *E. coli* host. This library was exposed to different stress conditions, including 6.7μg/ml kanamycin, 0.5μg/ml tetracycline, 2.0μg/ml chloramphenicol, 5.5% ethanol, pH=4.5, and pH=7.5. The enrichment and/or depletion of particular constructs were assessed through cloning site amplification of the pool plasmids and visualization on agarose gels. Competition assays were performed by mixing equal volumes of *E. coli* cells carrying tRNA-containing plasmids in a Δ*lacZ* background and *lacZ*$^+$ *E. coli* cells carrying empty plasmids, inoculating the stress-delivering culture with this mixture, and counting the red and white colony forming units (CFUs) on MacConkey plates after incubating overnight. Selection index R was defined as the ratio of logarithm of growth of tRNA-containing cells over logarithm of growth of cells carrying empty plasmids.

See Supplementary Methods for detailed description of mathematical and experimental procedures.

Software packages developed and used in this work are available online at http://webpages.mcgill.ca/staff/Group2/rsalav/web/Software/ICodPack/index.htm.

4.3 Results

4.3.1 A universal correlation between codon usage and function

We examined the relationship between codon usage and function in 785 organisms (including 72 eukaryotes, 661 bacteria and 52 archaea), the sequences and functions of whose genes were retrieved from Kyoto Encyclopedia of Genes and Genomes – KEGG [190]. Since paralogous duplicates usually have the same function as well as similar synonymous codon usages, their presence might result in over-estimating the similarity of codon usage among proteins of similar function. Therefore, duplicates were removed within each genome using nucleotide BLAST as described before [168]. In this work, we used a relatively large E-value cutoff of 0.001 to make sure that all duplicates were removed and the results are unbiased. For each gene, the usages of synonymous codons were calculated and normalized over the usages of their corresponding amino acids, here referred to as f_c. We applied suitable filters to reduce random fluctuations and obtain a robust measure of synonymous codon usage (see Supplementary Methods). The distance of a pair of genes i and j regarding the usage of codon c was calculated as $d_{ij}(c)=|f_{c,i}-f_{c,j}|$ [189]. If genes with similar functions use similar frequencies of synonymous codons, we shall expect a negative correlation between $d_{ij}(c)$ and the likelihood of sharing a biological function; i.e., the more dissimilar the synonymous codon usages of two genes, the less likely they participate in the same pathway. We observed significantly negative linear correlations between d and likelihood of functional linkage in almost all the examined genomes (see online Supplementary Table S1 at http://nar.oxfordjournals.org/content/suppl/2009/09/23/gkp792.DC1/nar-01747-s-2009-File014.doc), indicating a universal pattern in which genes of similar functions have similar usages of synonymous codons. Our set of genomes covered all taxonomic domains, although in particular negative correlations were highly significant in eukaryotes (Figure 4-1).

These results indicate that our previously observed pattern [189] in which functionally interacting proteins use similar codon usages is not restricted to a few organisms; rather, it is a universal characteristic of the genomes across all domains. We will investigate the cause of this pattern in the next section.

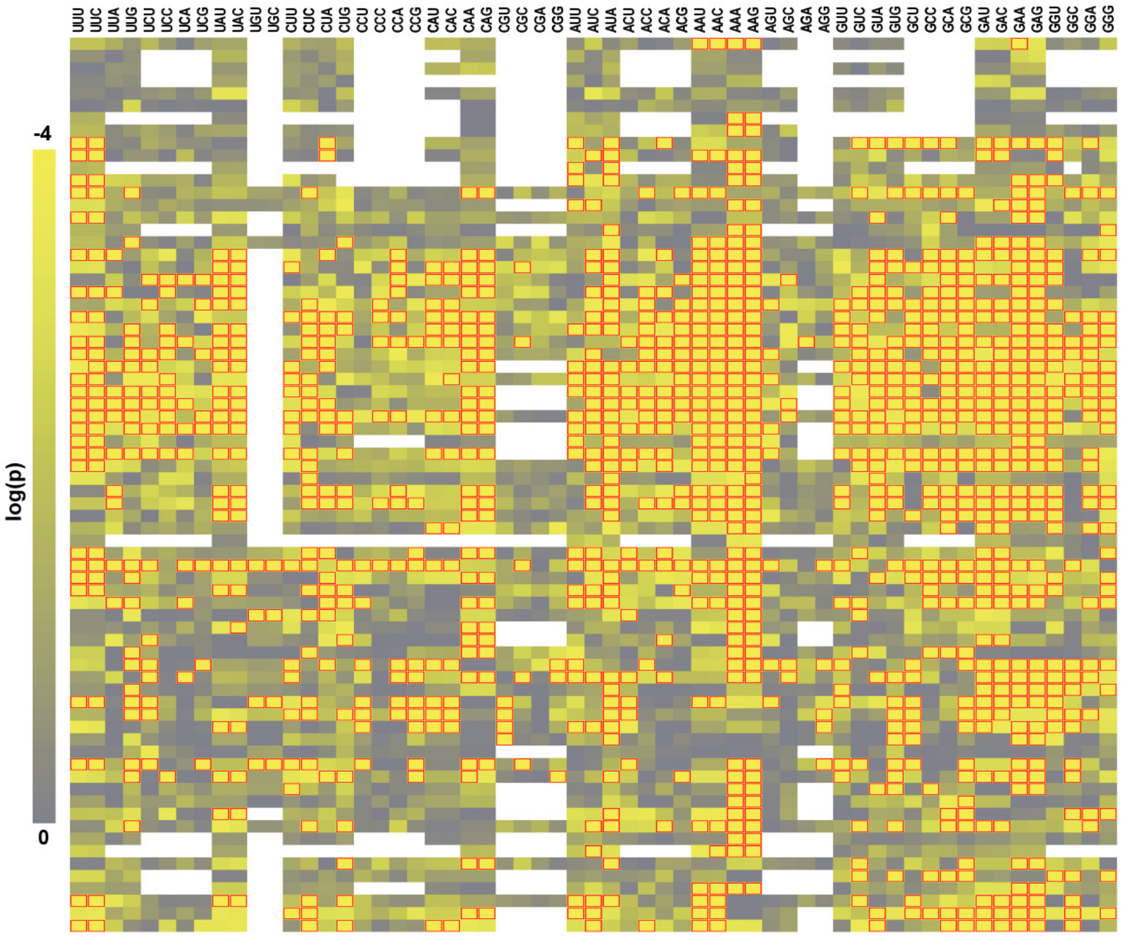


Figure 4-1. A heat map illustrating the significance of the negative correlations between *d* and likelihood of functional linkage in 72 eukaryotes and 59 codons – Each row represents one organism in the same order as the top 72 rows of online Supplementary Table S1 (http://nar.oxfordjournals.org/content/suppl/2009/09/23/gkp792.DC1/nar-01747-s-2009-File014.doc), while each column represents one codon. Stop codons, AUG and UGG are omitted. The *p*-values of the correlations are shown by a color gradient on log scale (left bar), with yellow color standing for small *p*-values. Significant correlations ($p \leq 1\times10^{-4}$) are highlighted by red frames, indicating that the corresponding codons are used similarly among the proteins that share the same function. The expected value of false discovery rate (FDR) is $<4\times10^{-4}$. White regions stand for cases in which the correlation coefficient could not be calculated due to lack of enough functional linkages.

4.3.2 Genes with similar expression patterns have similar synonymous codon usages

In the classic view on the relationship between codon usage and protein expression, a constant set of optimal codons is assumed for an organism over different life stages and conditions. This model implies that genes with a particular codon usage should have a translation efficiency that remains constant across different conditions. Assuming that this constant translation efficiency is selected based on the overall expression level of each gene [141], genes with similar codon usages should have similar "average" expression levels, but not necessarily similar expression "patterns". An alternative, unexplored hypothesis can be that the set of optimal codons is not constant, and changes from one condition to another. In this case, the translation efficiencies of genes that have similar codon usages do not remain constant, but change in a synchronized manner in response to the changes of the set of optimal codons. Thus, it is reasonable to assume that such genes would have similar expression "patterns".

Knowing that genes with similar functions have similar expression patterns ([191] and the references within), the observed similarity between codon usages of functionally linked proteins led us to reevaluate the two abovementioned hypotheses. We tested these hypotheses on four divergent organisms with available genome-wide expression profiles, human, yeast, *E. coli* and *C. elegans* [192-195]. In each organism, clusters of coexpressed genes (i.e., genes with similar expression patterns) were analyzed in the same way as we analyzed the clusters of functionally linked genes in the previous section. Similarly, we also clustered the genes in each organism according to their average expression level, and performed the same analysis. Figure 4-2 shows that in all of the tested genomes, codon usage has the strongest correlation with expression pattern rather than average expression level, corroborating the "variable set of optimal codons" hypothesis. Strikingly, this correlation is most obvious for the human genome, where most of the correlations are between -0.80 and -0.90 (online Supplementary Table S2 at http://nar.oxfordjournals.org/content/suppl/2009/09/23/gkp792.DC1/nar-01747-s-2009-File015.doc) and, with only one exception (correlation between CGU and coexpression), all of them are highly significant ($p \leq 1 \times 10^{-4}$).

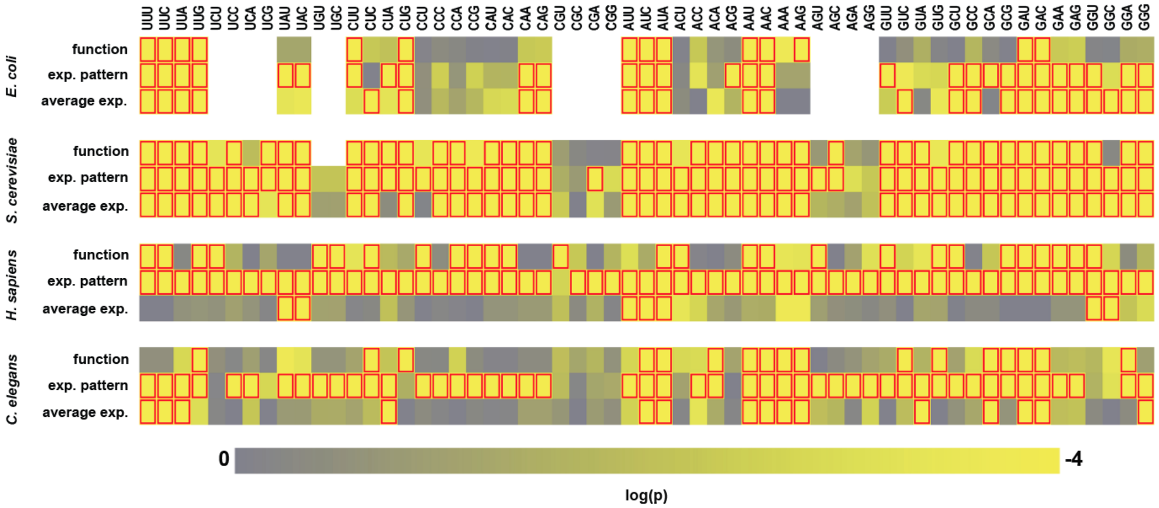


Figure 4-2. The significance of correlation between codon usage and clusters of genes according to different properties – Genes were clustered according to function, expression profile (resulting in clusters of coexpressed genes) or average gene expression level (resulting in clusters of genes with similar average expression levels). Functional clusters were obtained from KEGG pathway database [190]. Coexpression clusters for *S. cerevisiae*, *Homo sapiens* and *C. elegans* were derived from [168]. For *E. coli*, expression profiles of the genes were obtained from [194] and were clustered using Iclust [73]. Average gene expression levels were obtained by averaging the expression profile of each gene, except for *S. cerevisiae* where a previously reported reference mRNA level dataset was used [135]. The correlation between *d* and the likelihood of occurrence in the same cluster was assessed for each property in each organism for all codons, excluding stop codons, AUG and UGG. Significantly negative correlations are indicated by light red frames ($p \leq 1\times10^{-4}$). Refer to online Supplementary Table S2 (http://nar.oxfordjournals.org/content/suppl/2009/09/23/gkp792.DC1/nar-01747-s-2009-File015.doc) for the values associated with this figure.

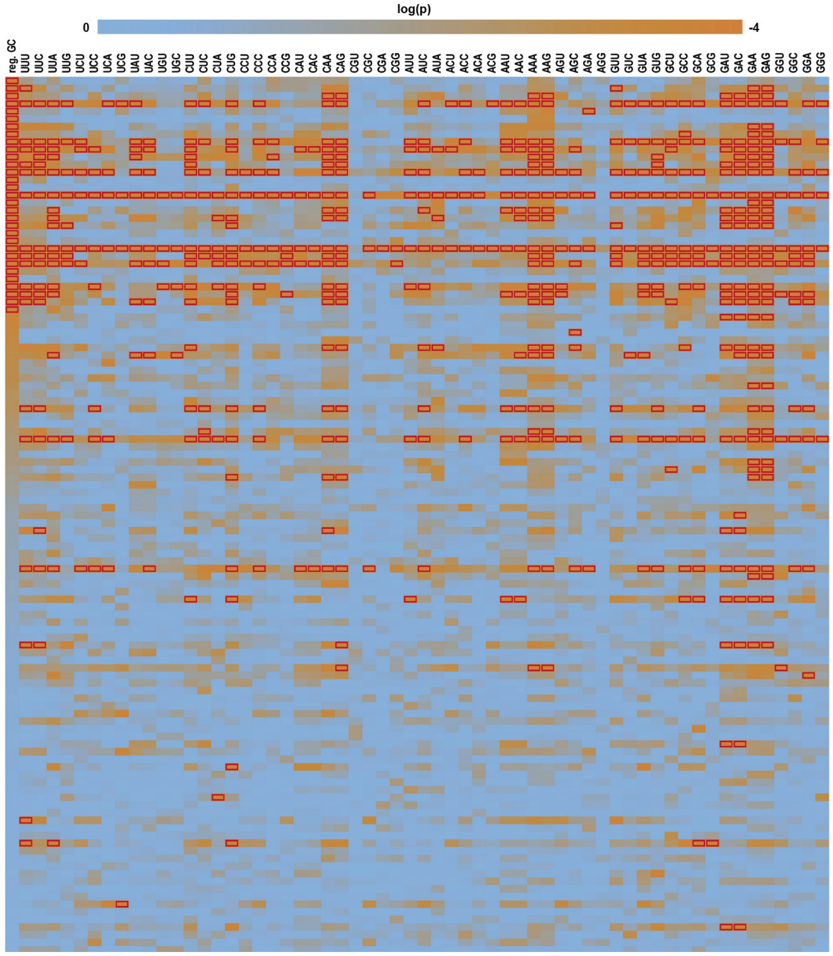


Figure 4-3. Mutual information of synonymous codon usage in human coexpression clusters is significantly higher than expected from a random distribution. Each row represents a coexpression cluster, while each column, except for the first column from left, represents a codon. Stop codons, AUG and UGG are omitted. Significant mutual information (MI) values are shown by red frames ($p \leq 1\times10^{-4}$). Therefore, a red frame around a square indicates that the genes within the corresponding coexpressoin cluster use similar frequencies of the corresponding codon. The expected value of false discovery rate (FDR) is 1×10^{-3}. MI of regional GC content in each coexpression cluster was assessed similarly (shown in the first column from left). Regional GC content was calculated as the GC content of the 50kb genomic region surrounding each gene, similar to [177]. Coexpression clusters are sorted according to the descending order of the MI of regional GC content; thus, the upper rows represent clusters whose genes have significantly similar regional GC contents, while the lower rows correspond to clusters whose genes occur in different GC contexts.

We further analyzed the codon usage of each coexpression cluster in human, following the same methodology that has been used before for finding informative regulatory elements [168]. We calculated the mutual information of the usage of each codon for each expression profile, and assessed whether the observed mutual information was significantly higher than expected by chance (see Supplementary Methods for details). A high mutual information value signifies a non-random usage of the corresponding codon among genes within the corresponding coexpression cluster. Figure 4-3 shows that, in many different coexpression clusters, synonymous codons are used non-randomly; in other words, specific frequencies of synonymous codons are preferred for each expression pattern, resulting in similar synonymous codon usages among the genes that are coregulated. Moreover, this non-random distribution of synonymous codon usage is not merely a result of similar GC content as reported before [177]. This is particularly obvious in coexpression clusters whose genes occur in genomic isochores with different GC contents but still show significantly similar synonymous codon usages (the lower half of Figure 4-3). It should be noted that non-random usage of a codon could be due to either preference for using that codon or preference for not using it (both resulting in high mutual information values). An example is shown in Supplementary Figure 4-1, where some coexpression clusters are over-represented among genes with high frequencies of codon UUU, while some other clusters are over-represented among genes that have low frequencies of UUU. Clustering the human genes based on their “average” expression level instead of expression pattern, we performed the same analysis and found no significant mutual information values. We also found no significant mutual information between expression pattern and the usage of any amino acid, indicating that the non-randomness of codon usage among coexpressed genes is independent of the amino acid context.

As a complementary method, we also clustered human genes just based on their synonymous codon usage (using 59 codons), and examined whether different coexpression groups show non-random distribution among these clusters. It is shown in Supplementary Figure 4-1 that many coexpression groups show significantly non-random distribution; each coexpression group is specifically enriched within certain codon usage clusters, while significantly under-represented in other clusters. A similar analysis on *S. cerevisiae* also reveals coexpressed genes with significantly similar synonymous codon usages. Interestingly, these coexpression clusters do not always consist of the most abundant genes. Instead, there exist many low-abundance coexpressed genes that show significant similarities regarding the usage of several synonymous codons (Supplementary Figure 4-2), although it is not as noticeable as in human.

4.3.3 Difference in codon usage directly affects regulation of protein expression

We examined whether modification of the codon usage of a gene can change the response of this gene to environmental conditions *in vivo*. To this end, we constructed a modified version of *lacZ* from *E. coli* K12-MG1655, in which the codon usage was changed considerably while keeping the original protein sequence (see Supplementary Methods). The original *lacZ* [GenBank:U00096, region 362455-365529] and the modified *lacZ* [GenBank:FJ839685] were cloned in yeast, and the expression pattern of LacZ was assessed in 16 different growth and stress conditions, using a quantitative β-galactosidase assay. Although the CAI values of these two genes were different (0.649 for modified *lacZ* compared to 0.213 for original *lacZ*), their average expression levels were not significantly different (paired Student's *t*-test score 0.86, $p<0.25$). However, as we expected, there were several conditions in which the protein expression was significantly different between the two variants. Particularly, three conditions yielded significantly higher galactosidase activities of the modified *lacZ* compared to the original *lacZ*, while two conditions yielded significantly higher activities of the original *lacZ* (Figure 4-4).

We propose that codon usage affects the regulation of protein expression by linking it to the regulation of tRNA composition in the cell. In other words, as in different conditions different proteins are required, tRNA composition of the cell may change accordingly in order to accommodate the changing demands for synthesis of new proteins. The response of genes to the new tRNA composition depends on their codon usages; hence, the translation efficiencies of genes with different codon usages change differently, causing the observed difference between the patterns of expression of the two *lacZ* variants. It has to be emphasized that this is a very likely, yet indirect conclusion from our experiment and we did not measure the tRNA content of yeast in the examined 16 conditions. In the next section, we hypothesize that not only the tRNA content may change according to the expression demands of the cell, but also we can change the cell phenotype by engineering the tRNA content.

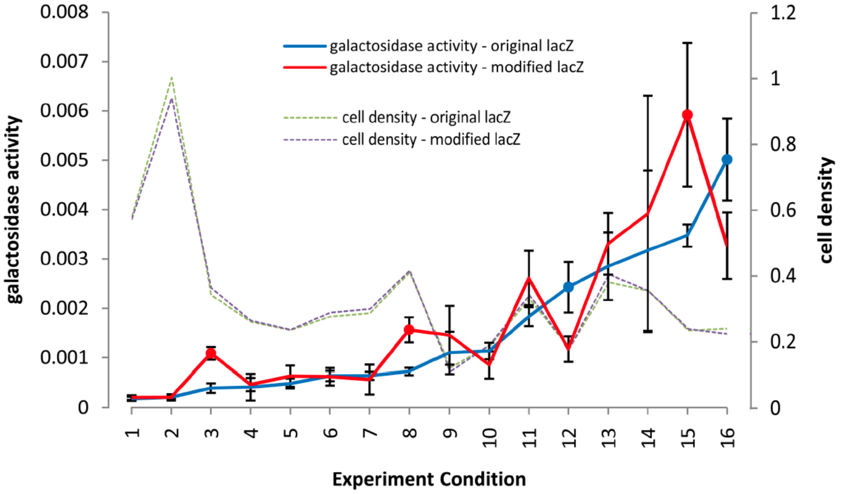


Figure 4-4. Expression pattern of *lacZ* gene varies as a result of codon usage modification – Two different *lacZ* sequences, i.e., the original gene from *E. coli* K12-MG1665 and a variant with modified codon usage, were expressed in yeast in different conditions, and galactosidase activity was measured. The two yeast strains carrying the two *lacZ* variants did not show any significant differences in their growth pattern (estimated by OD_{600}). However, the galactosidase activity was significantly different between the two strains in several conditions. The unit of galactosidase activity in this figure is $OD_{405}\times s^{-1}\times(OD_{600})^{-1}$. Blue circles indicate higher expression of original *lacZ*, while red circles indicate higher expression of modified *lacZ* ($p<0.05$ with Bonferroni correction for 16 experiments). Experiment conditions: (1) YPDA: 37°C; (2) YPDA: 30°C; (3) DTT shock: 30°C; (4) 2% sucrose: 37°C; (5) DTT shock: 37°C; (6) 2% ethanol: 37°C; (7) 2% glucose: 37°C; (8) 2% glucose: 30°C; (9) hyper-osmotic shock: 37°C; (10) SD: 37°C; (11) 2% ethanol: 30°C; (12) steady 1M sorbitol: 37°C; (13) 2% sucrose: 30°C; (14) SD: 30°C; (15) hyper-osmotic shock: 30°C; (16) steady 1M sorbitol: 30°C. Each experiment was performed in hexaplicate; the standard deviations are depicted by error bars.

4.3.4 Changes in tRNA abundance confer adaptative capacity

The results of the previous analyses and experiments suggest indirectly that the tRNA composition of the cell may follow its expression demands. But is tRNA composition also able to push the expression profile of the cell to a different state in order to cause a particular phenotype? We examined this by looking for phenotypic changes that might occur in the cell as a result of perturbation of its tRNA content.

Briefly, we constructed a plasmid library in pBAD18 backbone, each carrying a random selection of 25 tRNA genes from *E. coli* (see Supplementary Methods). From each tRNA gene, a plasmid could have zero, one, or multiple copies. Each copy could be oriented randomly in forward or reverse direction. The rational for this approach was that those sequences cloned in the forward orientation result in an overabundance of the tRNA, while those cloned in reverse most likely decrease the tRNA concentration through double strand formation with the tRNA transcribed in the cell. This library was transformed into *E. coli*, and the pool of transformed cells was grown under different environmental stresses. The initial frequency of constructs was visualized through cloning site amplification (Supplementary Figure 4-3). After one or two rounds of selection, the plasmid population was visualized to see whether certain constructs were enriched upon selection. We found two selection conditions in which particular plasmids had highly significant adaptive consequences in a short time-scale (~10 generations): sub-inhibitory concentrations of kanamycin (6.7μg/ml) and tetracycline (0.5μg/ml). In the case of kanamycin-containing medium, the enriched plasmid (named pBAD-tKAN) contained one copy of *glyT* tRNA gene in forward direction and one copy of *serW* tRNA gene in reverse direction. On the other hand, growth in the presence of tetracycline resulted in the enrichment of a plasmid containing *ileX* tRNA gene in forward direction, designated pBAD-tTET (in each case, 10 clones were randomly selected for sequence analysis; see Supplementary Methods). Repeating the experiment on the library resulted in selection of the same plasmids, indicating that the observed enrichment is selective and is not due to drift. We also confirmed that the selection of these particular tRNA isoacceptors was not a result of bias in the original library; all tRNA isoacceptors of Gly, Ser and Ile were represented in the library in both forward and reverse directions at different combinations (See Supplementary Methods and Supplementary Figure 4-3). This shows that, for example, in the presence of tetracycline, *ileX* has a significant fitness advantage over *ileT* since only *ileX* was selected.

Using a competition assay, we observed that in kanamycin-containing medium, *E. coli* cells freshly transformed with pBAD-tKAN have a selection index of about 1.5 over wild-type *E. coli* (MG1655) carrying an empty pBAD18 plasmid (*p*-value<0.025; for definition of selection index, see the Materials and Methods section). pBAD-tKAN confers no growth advantage over empty pBAD18 in tetracycline-containing medium (negative control). Similarly, pBAD-tTET-carrying cells with clean genomic background have a selection index of ~2 in tetracycline-containing medium (*p*-value<0.001), and no growth advantage in kanamycin-containing medium, indicating that each plasmid specifically increases the fitness in the medium at which it is selected.

4.4 Discussion

We showed that there is a universal correlation between codon usage and gene function, and that this correlation is even more obvious if we consider, instead of function, expression pattern as the basis for clustering the genes within each genome. The best hypothesis that can explain this observation as well as the results of our experiments is that the tRNA composition follows the expression demands of the cell. In other words, if in a particular condition a set of proteins with a particular function are needed and thus are expressed at high levels, tRNA composition changes accordingly to provide the required material. Since this adaptation would best work if in each condition the expressed genes had similar codon usages, a universal function-specificity has emerged in the codon usage within each genome.

The new hypothesis postulates that genes with similar expression patterns, even though having different average expression levels, should have similar codon usages. This is most obvious in organisms with complex developmental and physiological circuits such as human and *C. elegans* [196], in which there is a very strong correlation between codon usage and expression pattern but almost no correlation between codon usage and average expression level (Figure 4-2). In the case of simpler, fast-growing organisms such as yeast and *E. coli*, it is more difficult to discriminate between our new hypothesis and the conventional view, since there is a high correlation between average expression level and expression pattern: in these organisms, genes that have similar expression patterns usually show similar average expression levels as well. It is reasonable to think that in microorganisms with high growth rate, both the overall expression level of proteins and the dynamic pattern of expression may contribute to shaping the coding sequence. This is supported by the observation that in many cases the correlation coefficient of codon usage with expression pattern is still significant after correcting for the confounding effect of average expression level and vice versa (Supplementary Figure 4-4). However, the effect of expression pattern seems to be more profound than average expression level.

There have been previous reports suggesting that, via regulation of tRNA activity, genes with certain codon usages may be regulated in particular conditions [175, 197]. In this work, we showed that codon usage may have a wider effect on the response of a protein to environmental stimuli: in five out of 16 environmental conditions that we examined, a change in codon usage alters the extent and sometimes even the direction of LacZ response. Since both *lacZ* variants that we used had the same regulatory sequences in their upstream and downstream regions, the simplest explanation for the differences in their expression patterns is a difference in translation efficiency: while in some conditions the original *lacZ* showed greater translation efficiency, in some other conditions the modified *lacZ* exhibited greater translation efficiency. This corroborates the "variable optimal codon set" hypothesis. This is not however the only explanation; for example, the translation efficiency of LacZ may be affected by, in addition to codon usage, the structure of its mRNA. To examine the latter case, we studied the free folding energy for the critical regions of mRNAs of the two *lacZ* variants, and found no significant differences (Supplementary Figure 4-5). Our results are congruent with a recent report that the folding energy affects overall expression level [187]: indeed the average expression levels of the two *lacZ* variants were similar; rather, the expression patterns were different. To our knowledge, there is no known mechanism based on which mRNA structure, without involvement of regulated *trans*-acting factors, could affect the expression pattern.

We also showed that changes in tRNA composition may bring about significant adaptive consequences, such as higher resistance to particular antibiotics. This means that changes in tRNA composition results in tangible phenotypic effects, thus suggesting the possibility that tRNA composition not only follows the expression demands of the cell, but may also change the expression profile of the cell on its own.

The antibiotics that we examined suppress cell growth by inhibition of translation. Thus, it might be argued that the plasmids pBAD-tKAN and pBAD-tTET confer resistance to these antibiotics by overexpression of tRNAs and, hence, generally enhancing translation. However, this scenario seems unlikely due to the nature of the tRNAs that these plasmids carry: both *glyT* and *ileX* encode tRNAs that recognize rare codons in *E. coli* [198] (GGA/G and AUA, respectively). This is while the overall rate of translation would be enhanced much more efficiently if tRNAs that could recognize abundant codons were overexpressed. In fact, overexpression of *glyT* and *ileX* has no direct effect on translation efficiency of many highly demanded genes in *E. coli*, as these genes lack the cognate codons of these tRNAs. Furthermore, selection of the reverse complement of *serW* cannot be explained by enhancement of translation: the reverse complement of *serW* is assumed to inhibit Seryl-tRNA 5 through direct binding. Specificity of pBAD-tKAN and pBAD-tTET in conferring resistance towards only kanamycin and tetracycline, respectively, and not vice versa, adds to the above reasons to conclude that the mechanism of action of these constructs is not through a general enhancement of translation.

It is interesting to see that the combination of *glyT* in forward direction (*glyT$_f$*) and *serW* in reverse direction (*serW$_r$*), and not *glyT$_f$* or *serW$_r$* alone, was selected in the presence of kanamycin. This indicates that the combination of these two is much stronger in conferring kanamycin resistance than any of them alone. Assuming that such cooperative action of tRNAs can have beneficial effects on cell fitness in other situations as well, a regulatory network that manages and synchronizes the activity of different tRNAs can be readily hypothesized. This also suggests that usages of different codons may have coevolved.

Although in this experiment we tested the effect of tRNA concentration on phenotype, cells may potentially change the activity profiles of tRNAs in ways other than changing the concentration as well. For example, enzymatic modification of tRNA has been shown to change the codon preference [175]. This mechanism especially seems possible since even for amino acids that have only two codons there is an expression pattern-specific codon usage (Figure 4-2 and Figure 4-3). In most organisms, there is only one kind of tRNA for each of the amino acids with two U/C-ending codons. It is thus not possible to change which codon is preferred by varying the concentration of this tRNA, as this tRNA recognizes both of the codons. However, enzymatic modification of tRNA can result in a change in its preference towards its two cognate codons [175], which can describe expression pattern-specificity of codon usage for two-codon amino acids. This indicates that variation of tRNA composition may extend beyond concentration and may include variation of tRNA activity by means such as enzymatic modification as well.

The above experiments suggest a novel approach for modulating functions with applications in biology and biotechnology: we showed that perturbations to a single tRNA gene may change the survival of the cell in certain conditions. It can also be anticipated that certain metabolic pathways will be enhanced by overexpression of tRNAs that recognize the specific codons of those pathways. Also, pathway engineering may benefit from more careful design of coding sequences regarding their codon usage.

Last but not least, the above observations lead to a novel method for prediction of gene expression profiles and gene functions. We employed a naïve Bayesian network to construct a classifier that is able to predict expression profile or the function of a gene solely based on its codon usage. Using this classifier, we were able to achieve a high sensitivity and specificity in predicting many human coexpression clusters. Furthermore, applying the same classifier on functional groups instead of coexpression clusters showed that some functions can be reliably predicted based on synonymous codon usage (Supplementary Figure 4-6). We anticipate that this method can considerably enhance homology-independent annotation of genes, especially for genomes whose genes are not conserved in well-studied model organisms (see Supplementary Figure 4-6 for a few examples in *Trypanosoma brucei*, the causative agent of human African trypanosomiasis).

4.5 Supplementary Methods

4.5.1 Genome sequences, functional annotations and coexpression clusters

All sequences and gene-pathway assignments were retrieved from KEGG GENES database [190] on Aug 1, 2008. In each genome, paralogues were recognized using nucleotide BLAST with the relatively large E-value cutoff of 0.001 and removed in order to avoid any biases due to presence of duplicate genes. Coexpression clusters for human, mouse, yeast and *Caenorhabditis elegans* were the same as used before by Elemento *et al.* [168]. Elemento *et al.* used Iclust [73] to cluster the genes based on their previously reported expression profiles over different tissues [192], different environmental conditions [193], and/or different developmental stages and varieties of mutants [195]. We obtained coexpression clusters for *Escherichia coli* genes by applying Iclust [73] on a previously compiled compendium of *E. coli* gene expression profiles [194].

4.5.2 Calculating the normalized frequency of each codon

The normalized frequency of each codon in each gene, referred in this article as f_c, was calculated as the usage of that codon divided by the usage of the amino acid it codes for:

$$f_c = \frac{n(c)}{n(a_c)}$$

Therefore, f_c is a measure of synonymous codon usage and does not reflect amino acid usages (f_c is directly proportional to RSCU [147]). For each gene, we calculated this value for codons of an amino acid only if the frequency of that amino acid was at least ζ times the number of its codons, i.e., the average expected number of each codon could not be less than ζ, assuming a uniform distribution of synonymous codons. For example, when ζ=5, if phenyl alanine was used less than 10 times in a gene, we did not calculate f_{UUU} and f_{UUC} for that gene. This filtering reduces the random fluctuations in f_c and provides a robust measure of codon usage. Wherever an amino acid did not meet the above criteria, f_c values of its codons were treated as missing data. For the analysis of the genomes of human and *C. elegans*, ζ was set to 10. For all other analyses, ζ was set to 8.

4.5.3 Analysis of the correlation coefficient between `codon usage distance and co-clustering

The distance of a pair of genes *i* and *j* regarding their usage of codon *c* was calculated as $d_{ij}(c)=|f_{c,i}-f_{c,j}|$ [189]. The distances of all gene pairs were calculated, and gene pairs were sorted according to their *d* values. Then, the sorted gene pairs were divided into several equally populated bins, and for each bin b_i, the likelihood of being in the same cluster (either functional cluster or cluster of coregulated genes) was calculated as:

$$L_i = p(b_i|pos)/p(b_i|neg) = \frac{n(b_i \bigcap pos)/n(pos)}{n(b_i \bigcap neg)/n(neg)}$$

In the above equation, *pos* means being co-clustered, and *neg* means not being co-clustered. Two genes were considered co-clustered if there was at least one cluster containing both of them. Pearson correlation coefficient between the minimum *d* value of each bin and the *L* value associated with that bin was calculated. Significance of Pearson correlation coefficient was estimated by randomly shuffling gene-cluster assignments 10^4 times, each time repeating the calculations and comparing them with the original correlation coefficient.

4.5.4 Analysis of the mutual information between a continuous property of genes and a particular gene cluster

Given a variable γ and a particular gene cluster α, we are interested to know whether the distribution of γ in α is random or not. We use mutual information (MI) to capture such non-random relationships. The variable γ can be any continuous or semi-continuous property of genes such as synonymous codon usage, regional GC content, genomic location, etc. The cluster α can represent a set of coregulated genes, a set of genes that participate in the same metabolic pathway, etc. The set of genes that are not in cluster α is called α′. The genes in α+α′ are sorted according to the value of γ, and are divided into *m* equally populated bins. A 2×*m* table is formed in which the element $e_{1,i}$ shows the number of genes in the *i*th bin that are in α, and the element $e_{2,i}$ shows the number of genes in the *i*th bin that are in α′ ($1 \leq i \leq m$). Then, the value of MI across this table is calculated as described before [168]. To examine whether the obtained MI is significantly higher than would be expected from a random distribution, the gene-cluster assignments are randomly shuffled *n* times, MI is calculated each time, and the probability of observing a random MI ≥ the original MI is calculated.

In this work, *m* (the number of bins) was set to 5 for the analysis of codon usage, and10 for the analysis of regional GC content. Gene-cluster assignments were shuffled 10^4 times for the assessment of significance of MI.

4.5.5 Prediction of gene-cluster assignments by means of codon usage

We used a naïve Bayesian network to predict gene-cluster assignments based on codon usage. Naïve Bayesian networks assume that the properties based on which they classify the objects are independent. Thus, the likelihood that the gene *g* belongs to cluster α is calculated as:

$$L(g \in \alpha \mid C) = \prod_{c \in C} L(g \in \alpha \mid f_c)$$

The cluster α can be a coexpression cluster, a metabolic pathway, etc. Here, *C* is the set of codons that are used for classification of *g*, $L(g \in \alpha \mid C)$ is the likelihood that *g* belongs to α given the set of codons *C*, and $L(g \in \alpha \mid f_c)$ represents the likelihood that *g* belongs to α given its normalized usage of codon *c*, i.e., f_c. If f_c is missing, $L(g \in \alpha \mid f_c) = 1$. For more information about the calculation of conditional likelihoods and implementation of naïve Bayesian networks, refer to [81].

The degree of freedom for synonymous codon usage of each amino acid is always one unit less than the number of synonymous codons of that amino acid. Therefore, ignoring one codon from each amino acid, we will still have the same amount of information. *C* is chosen in a way to avoid such redundancies as well as to maximize the prediction power. Briefly, codons are selected iteratively, starting from the one that can best distinguish between α and α′. Then, at each iteration, all codons are tested and the one whose addition to *C* results in the maximum prediction power is selected. Prediction power is assessed by the area under the ROC curve (Receiver Operating Characteristic or ROC curve plots sensitivity against false positive discovery rate). This procedure is repeated until *C* contains a predefined number of codons. Since stop codons are excluded and one codon from each amino acid is ignored, the maximum possible size of *C* is 41.

Software packages developed and used in this work are available online at http://webpages.mcgill.ca/staff/Group2/rsalav/web/Software/ICodPack/index.htm.

4.5.6 Investigating the effect of changing codon usage on expression pattern of lacZ

We have hypothesized that codon usage regulates protein expression by responding to tRNA composition dynamics. One of the most direct implications of this hypothesis is that changing the codon usage of a gene will change the way it responds to alterations of tRNA composition. We examined two variants of a gene with essentially the same protein sequence but different codon usages in different growth and stress conditions, with the assumption that these different conditions could induce alteration of tRNA composition. We selected *lacZ* from *Escherichia coli* K12-MG1655 (GenBank U00096, region 362455-365529) for this purpose. The activity of this gene can easily be monitored in a β-galactosidase assay. Furthermore, yeast does not contain a homolog of this gene, making it possible to measure the activity of only the ectopic allele. Also, the relatively long sequence of this gene (1024 codons) makes it possible to design variants with desired codon usages.

The coding sequence of our modified *lacZ* (GenBank FJ839685) in alignment with the original *lacZ* can be found in Appendix I. The modified *lacZ* was synthesized by GenScript (New Jersey). The original *lacZ* was amplified from the genomic DNA of *E. coli*. The two variants of *lacZ* were cloned into pBridge (Clontech #630404) using *NotI* and *BglII* restriction sites. AH109 yeast cells (Clontech #630444) were then transformed by the two constructs according to YeastMaker™ Yeast Transformation System 2 (Clontech #630439) User Manual (document PT1172-1). Six clones for each construct were chosen randomly and used for inoculating 5ml of Minimal SD supplemented with –

Met/–Trp DO Supplement (Clontech #630431). After 72h incubation at 30°C (250rpm), OD_{600} of each culture was adjusted to 0.5 by adding suitable amount of medium, and 25µl was used to inoculate 500µl of eight different culture media, as follows (the base for all media is –Met/–Trp DO-supplemented SD, except for YPDA) :

(A) –Met/–Trp DO-supplemented SD; (B) YPDA (Clontech #630306); (C) 2% ethanol; (D) 2% glucose; (E) 2% sucrose; (F) –Met/–Trp DO-supplemented SD; (G) 1M sorbitol; (H) –Met/–Trp DO-supplemented SD.

From each medium, two 500µl aliquots were inoculated, one of which was incubated at 30°C and the other one at 37°C (overnight, 250rpm). A total of 192 cultures were prepared (8 media × 2 temperatures × 2 yeast transformants × 6 replicates). After 16h, an additional 500µl of each medium was added to the corresponding cultures, except for media F and H, to which 500µl of –Met/–Trp DO-supplemented SD containing either 5mM DTT or 2M sorbitol (for hyper-osmotic shock) was added, respectively. Cultures were incubated again for 120min, after which β-galactosidase activity was measured as follows:

OD_{600} of each culture was determined. Cells were spun down at 5000rpm for 2min and supernatants were decanted. Pellets were shock-frozen in liquid nitrogen, thawed at room temperature water bath, and resuspended in 153µl Buffer H (100ml HEPES, 150mM NaCl, 2mM $MgCl_2$, 1% BSA). To each suspension, 11µl 0.1% SDS and 11µl chloroform was added and vortexed for 1min at the highest settings. Then, from each extract, 140µl was transferred to assay plate, where 20µl of 4mg/ml ortho-Nitrophenyl-β-galactoside (ONPG, Sigma #N1127) solution was added. OD_{405} was monitored for 25min, and the slope at the beginning of the resulting curve was considered as the β-galactosidase activity. This activity was divided by OD_{600} of cultures in order to normalize for cell density.

4.5.7 In vivo selection of a tRNA library for resistance against stress conditions

We hypothesized that for each environmental condition there should be an optimal tRNA composition that confers the highest fitness to the organism specifically for that condition, based on the following premises:

1. At each environmental condition, specific functions need to be up-regulated and certain functions need to be down-regulated.

2. We see a function-specific codon usage among genes.

3. tRNA composition affects translation rate of genes with different codon usages differently. Thus, at each condition, a tRNA composition is optimal if it results in higher translation rate for functions that are demanded at that condition, and lower translation rate for functions that are not needed.

We selected a subset of *E. coli* tRNAs so that (i) each codon was represented by at most one tRNA and (ii) each tRNA had at least one iso-accepting partner in this subset (i.e., another tRNA that recognized a different codon, but corresponded to the same amino acid). A total of 25 tRNAs were selected this way. Primers were designed so as to amplify these tRNA genes along with about 40bp upstream and 40bp downstream of them (Appendix II). Amplified fragments were pooled and cloned into pBAD18 using *XmaI* restriction site and electroporated into *E. coli* K12-MG1655Δ*lacZ* host. LB+Amp was used for culture and selection of transformants.

For selection under each of the stress conditions, pool of transformants was cultured in the corresponding medium for two passages (100X dilution after each passage), after each of which the result of amplification of the cloning site in the pool of plasmids was analyzed on agarose gel (see Supplementary Figure 4-3 for some examples). In case of kanamycin and tetracycline, where selection of particular plasmids could be observed after the second round of selection, the sequences of 10 clones were analyzed (Appendix III and Appendix IV).

The selected plasmids contained *glyT*, *ileX* and *serW* (see the article for more details). In order to show that other isoacceptors of Gly, Ile and Ser were also present in the original library although they were not selected, we used our library as the template in individual

PCRs for *glyT*, *glyU*, *glyV*, *ileT*, *ileX*, *serU*, *serV* and *serW*. Each gene was amplified from the library using the primer CCATAGCATTTTTATCCATAAG, which binds to the upstream region of pBAD18 cloning site, and either the forward or reverse tRNA-specific primer (Appendix II). All of these tRNA genes were present in the library in both forward and reverse directions at different combinations (Supplementary Figure 4-3), indicating that selection of *glyT*, *ileX* and *serW* was fitness-directed rather than the result of biased starting library.

For assessing the fitness of the selected clones in comparison to cells carrying no ectopic tRNA alleles, 1μl from overnight culture of each strain was mixed with 1μl from overnight culture of *E. coli* cells carrying empty pBAD18 plasmid, in either LB+Amp+Kan or LB+Amp+Tet. After an overnight incubation, diluted cultures were spread on LB+Amp+X-gal+IPTG plates, incubated again overnight, and colonies were counted (colonies having empty pBAD18 are in a $lacZ^+$ background and are distinguishable from colonies having pBAD18-tRNA by their red color on MacConkey plates). Selection index R is defined as:

$$R = \frac{\log\left[n\left(pBAD18 \cdot tRNA_{final}\right)/n\left(pBAD18 \cdot tRNA_{initial}\right)\right]}{\log\left[n\left(pBAD18_{final}\right)/n\left(pBAD18_{initial}\right)\right]}$$

In the above equation, $n(\mathrm{pX}_{\mathrm{intial}})$ and $n(\mathrm{pX}_{\mathrm{final}})$ are the counts of the cells carrying pX before and after the overnight competition with the other strain, respectively. If pBAD18-tRNA-carrying cells have a growth advantage over wild-type cells, R index >1 is expected.

4.5.8 Appendix I

Alignment of original *lacZ* from *E. coli* K12-MG1655 (U00096, region 362455-365529) and our modified *lacZ*

```
original.lacZ      ATGACCATGATTACGGATTCACTGGCCGTCGTTTTACAACGTCGTGACTGGGAAAACCCT
modified.lacZ      ATGACCATGATCACCGACTCTTTGGCCGTCGTCTTGCAACGGAGAGACTGGGAAAACCCA
                   *********** ** ** **  ********** ** *****  * ************** 
```

```
original.lacZ      GGCGTTACCCAACTTAATCGCCTTGCAGCACATCCCCCTTTCGCCAGCTGGCGTAATAGC
modified.lacZ      GGTGTAACCCAATTGAACAGGTTGGCCGCCCACCCACCATTCGCCTCTTGGAGAAACTCT
                   ** ** ****** * **  *  * ** ** ** ** ** ******   *** * **


original.lacZ      GAAGAGGCCCGCACCGATCGCCCTTCCCAACAGTTGCGCAGCCTGAATGGCGAATGGCGC
modified.lacZ      GAAGAAGCCAGGACCGACAGGCCATCTCAACAACTTCGGTCTTTGAACGGTGAATGGAGG
                   ***** *** * *****  * ** ** *****  * **     **** ** ****** *


original.lacZ      TTTGCCTGGTTTCCGGCACCAGAAGCGGTGCCGGAAAGCTGGCTGGAGTGCGATCTTCCT
modified.lacZ      TTCGCCTGGTTCCCAGCCCCAGAAGCCGTACCAGAAAGCTGGTTGGAATGCGACTTGCCA
                   ** ******** ** ** ******** ** ** ********* **** *****  * **


original.lacZ      GAGGCCGATACTGTCGTCGTCCCCTCAAACTGGCAGATGCACGGTTACGATGCGCCCATC
modified.lacZ      GAAGCCGACACCGTCGTCGTACCCTCTAACTGGCAAATGCACGGTTACGACGCCCCAATC
                   ** ***** ** ******** ***** ******** ************** ** ** ***


original.lacZ      TACACCAACGTGACCTATCCCATTACGGTCAATCCGCCGTTTGTTCCCACGGAGAATCCG
modified.lacZ      TACACCAACGTCACCTACCCAATCACCGTCAACCCACCATTCGTCCCAACCGAAAACCCA
                   *********** ***** ** ** ** ***** ** ** ** ** ** ** ** ** **


original.lacZ      ACGGGTTGTTACTCGCTCACATTTAATGTTGATGAAAGCTGGCTACAGGAAGGCCAGACG
modified.lacZ      ACCGGTTGCTACTCTTTGACCTTCAACGTCGACGAATCTTGGTTGCAAGAAGGTCAAACC
                   ** ***** *****  * ** ** ** ** ** ***   *** * ** ***** ** **


original.lacZ      CGAATTATTTTTGATGGCGTTAACTCGGCGTTTCATCTGTGGTGCAACGGGCGCTGGGTC
modified.lacZ      AGGATCATCTTCGACGGCGTCAACTCGGCCTTCCACTTGTGGTGTAACGGTAGATGGGTC
                    * ** ** ** ** ***** ******** ** **  ******* *****  * ******


original.lacZ      GGTTACGGCCAGGACAGTCGTTTGCCGTCTGAATTTGACCTGAGCGCATTTTTACGCGCC
modified.lacZ      GGTTACGGTCAAGACTCGAGACTTCCAAGCGAATTCGACTTGTCGGCCTTCTTGAGGGCC
                   ******** ** ***    *  * **    ***** *** **   ** ** **  * ***


original.lacZ      GGAGAAAACCGCCTCGCGGTGATGGTGCTGCGCTGGAGTGACGGCAGTTATCTGGAAGAT
modified.lacZ      GGAGAAAACCGATTGGCCGTCATGGTCTTGAGATGGTCGGACGGCTCTTACTTGGAAGAC
                   ***********  * ** ** *****  ** * ***   ******  ***  *******


original.lacZ      CAGGATATGTGGCGGATGAGCGGCATTTTCCGTGACGTCTCGTTGCTGCATAAACCGACT
modified.lacZ      CAAGACATGTGGAGGATGTCTGGTATCTTCAGGGACGTCTCTTTGTTGCACAAGCCAACC
                   ** ** ****** *****   ** ** *** * ******** *** **** ** ** **


original.lacZ      ACACAAATCAGCGATTTCCATGTTGCCACTCGCTTTAATGATGATTTCAGCCGCGCTGTA
modified.lacZ      ACCCAAATCAGCGACTTCCACGTCGCCACCAGATTCAACGACGACTTCTCTAGGGCCGTC
                   ** *********** ***** ** *****  * ** ** ** ** ***    * ** **


original.lacZ      CTGGAGGCTGAAGTTCAGATGTGCGGCGAGTTGCGTGACTACCTACGGGTAACAGTTTCT
modified.lacZ      TTGGAAGCCGAAGTCCAAATGTGTGGTGAATTGCGGGACTACTTGCGAGTCACCGTCTCT
                    **** ** ***** ** ***** ** ** ***** ****** * ** ** ** ** ***
```

```
original.lacZ      TTATGGCAGGGTGAAACGCAGGTCGCCAGCGGCACCGCGCCTTTCGGCGGTGAAATTATC
modified.lacZ      TTGTGGCAAGGTGAAACCCAAGTCGCCTCTGGAACCGCCCCATTCGGTGGTGAAATCATC
                   ** ***** ******** ** ******   ** ***** ** ***** ******** ***


original.lacZ      GATGAGCGTGGTGGTTATGCCGATCGCGTCACACTACGTCTGAACGTCGAAAACCCGAAA
modified.lacZ      GACGAACGGGGTGGTTACGCCGACAGAGTCACCTTGAGGTTGAACGTCGAAAACCCAAAG
                   ** ** ** ******** *****  * *****  *  *  **************** **


original.lacZ      CTGTGGAGCGCCGAAATCCCGAATCTCTATCGTGCGGTGGTTGAACTGCACACCGCCGAC
modified.lacZ      TTGTGGTCTGCCGAAATCCCAAACTTGTACAGAGCCGTCGTCGAATTGCACACCGCCGAC
                    *****   *********** **  * **  * ** ** ** *** **************


original.lacZ      GGCACGCTGATTGAAGCAGAAGCCTGCGATGTCGGTTTCCGCGAGGTGCGGATTGAAAAT
modified.lacZ      GGTACCTTGATCGAAGCCGAAGCCTGCGACGTCGGTTTCAGAGAAGTCCGGATCGAAAAC
                   ** **  **** ***** *********** ********* * ** ** ***** *****


original.lacZ      GGTCTGCTGCTGCTGAACGGCAAGCCGTTGCTGATTCGAGGCGTTAACCGTCACGAGCAT
modified.lacZ      GGTTTGTTGTTGTTGAACGGTAAGCCATTGTTGATCAGGGGTGTCAACAGGCACGAACAC
                   *** ** ** ** ******* ***** *** ****  * ** ** *** * ***** **


original.lacZ      CATCCTCTGCATGGTCAGGTCATGGATGAGCAGACGATGGTGCAGGATATCCTGCTGATG
modified.lacZ      CACCCATTGCACGGTCAAGTCATGGACGAACAAACCATGGTCCAAGACATCTTGTTGATG
                   ** **  **** ***** ******** ** ** ** ***** ** ** *** ** *****


original.lacZ      AAGCAGAACAACTTTAACGCCGTGCGCTGTTCGCATTATCCGAACCATCCGCTGTGGTAC
modified.lacZ      AAGCAAAACAACTTCAACGCCGTCCGATGCTCTCACTACCCAAACCACCCATTGTGGTAC
                   ***** ******** ******** ** ** ** ** ** ** ***** **  ********


original.lacZ      ACGCTGTGCGACCGCTACGGCCTGTATGTGGTGGATGAAGCCAATATTGAAACCCACGGC
modified.lacZ      ACCTTGTGCGACAGATACGGTTTGTACGTCGTCGACGAAGCCAACATCGAAACCCACGGT
                   **  ******** * *****  **** ** ** ** ******** ** ***********


original.lacZ      ATGGTGCCAATGAATCGTCTGACCGATGATCCGCGCTGGCTACCGGCGATGAGCGAACGC
modified.lacZ      ATGGTCCCAATGAACCGGTTGACCGACGACCCACGCTGGTTGCCAGCCATGTCTGAAAGG
                   ***** ******** **  ******* ** ** ****** * ** ** ***   *** *


original.lacZ      GTAACGCGAATGGTGCAGCGCGATCGTAATCACCCGAGTGTGATCATCTGGTCGCTGGGG
modified.lacZ      GTCACCAGGATGGTCCAAAGGGACCGGAACCACCCATCTGTCATCATCTGGTCTTTGGGT
                   ** **  * ***** **  * ** ** ** *****   *** ***********  ****


original.lacZ      AATGAATCAGGCCACGGCGCTAATCACGACGCGCTGTATCGCTGGATCAAATCTGTCGAT
modified.lacZ      AACGAATCTGGTCACGGAGCCAACCACGACGCCTTGTACAGGTGGATCAAGTCTGTCGAC
                   ** ***** ** ***** ** ** ********  ****  * ******** ********


original.lacZ      CCTTCCCGCCCGGTGCAGTATGAAGGCGGCGGAGCCGACACCACGGCCACCGATATTATT
modified.lacZ      CCATCTAGGCCAGTCCAATACGAAGGTGGTGGAGCCGACACCACCGCCACCGACATCATC
                   ** **  * ** ** ** ** ***** ** ************** ******** ** **
```

```
original.lacZ      TGCCCGATGTACGCGCGCGTGGATGAAGACCAGCCCTTCCCGGCTGTGCCGAAATGGTCC
modified.lacZ      TGCCCAATGTACGCCCGAGTCGACGAAGACCAACCATTCCCAGCCGTCCCAAAGTGGTCT
                   ***** ******** ** ** ** ******** ** ***** ** ** ** ** *****


original.lacZ      ATCAAAAAATGGCTTTCGCTACCTGGAGAGACGCGCCCGCTGATCCTTTGCGAATACGCC
modified.lacZ      ATCAAGAAGTGGCTTTCTTTGCCAGGTGAAACCAGGCCATTGATCTTGTGCGAATACGCC
                   ***** ** ********  * ** ** ** **  * **  ***** * ************


original.lacZ      CACGCGATGGGTAACAGTCTTGGCGGTTTCGCTAAATACTGGCAGGCGTTTCGTCAGTAT
modified.lacZ      CACGCCATGGGTAACTCTCTTGGAGGTTTCGCCAAGTACTGGCAAGCCTTCAGGCAATAC
                   ***** *********  ****** ******** ** ******** ** **  * ** **


original.lacZ      CCCCGTTTACAGGGCGGCTTCGTCTGGGACTGGGTGGATCAGTCGCTGATTAAATATGAT
modified.lacZ      CCACGGCTTCAAGGTGGTTTCGTCTGGGACTGGGTCGACCAATCTTTGATCAAGTACGAC
                   ** **  * ** ** ** ***************** ** ** **  **** ** ** **


original.lacZ      GAAAACGGCAACCCGTGGTCGGCTTACGGCGGTGATTTTGGCGATACGCCGAACGATCGC
modified.lacZ      GAAAACGGTAACCCATGGTCTGCCTACGGTGGTGACTTCGGCGACACCCCAAACGACAGA
                   ******** ***** ***** ** ***** ***** ** ***** ** ** *****  *


original.lacZ      CAGTTCTGTATGAACGGTCTGGTCTTTGCCGACCGCACGCCGCATCCAGCGCTGACGGAA
modified.lacZ      CAATTCTGCATGAACGGTTTGGTCTTCGCCGACAGGACCCCACACCCAGCCTTGACCGAA
                   ** ***** ********* ******* ****** * ** ** ** *****  **** ***


original.lacZ      GCAAAACACCAGCAGCAGTTTTTCCAGTTCCGTTTATCCGGGCAAACCATCGAAGTGACC
modified.lacZ      GCCAAGCACCAACAACAATTCTTCCAATTCAGGTTGTCTGGTCAAACCATCGAAGTCACC
                   ** ** ***** ** ** ** ***** *** * ** ** ** ************** ***


original.lacZ      AGCGAATACCTGTTCCGTCATAGCGATAACGAGCTCCTGCACTGGATGGTGGCGCTGGAT
modified.lacZ      TCTGAATACTTGTTCAGACACTCTGACAACGAATTGTTGCACTGGATGGTCGCCTTGGAC
                      ****** ***** * **    ** *****  *  ************* **  ****


original.lacZ      GGTAAGCCGCTGGCAAGCGGTGAAGTGCCTCTGGATGTCGCTCCACAAGGTAAACAGTTG
modified.lacZ      GGTAAGCCATTGGCCTCTGGTGAAGTCCCATTGGACGTCGCCCCACAAGGTAAGCAATTG
                   ********  ****    ******** **  **** ***** *********** ** ***


original.lacZ      ATTGAACTGCCTGAACTACCGCAGCCGGAGAGCGCCGGGCAACTCTGGCTCACAGTACGC
modified.lacZ      ATCGAATTGCCAGAATTGCCACAACCAGAATCTGCCGGTCAATTGTGGTTGACCGTCCGA
                   ** *** **** *** * ** ** ** **    ***** *** * *** * ** ** **


original.lacZ      GTAGTGCAACCGAACGCGACCGCATGGTCAGAAGCCGGGCACATCAGCGCCTGGCAGCAG
modified.lacZ      GTCGTCCAACCAAACGCCACCGCCTGGTCTGAAGCCGGTCACATCTCTGCCTGGCAACAA
                   ** ** ***** ***** ***** ***** ******** ******   ******** **


original.lacZ      TGGCGTCTGGCGGAAAACCTCAGTGTGACGCTCCCCGCCGCGTCCCACGCCATCCCGCAT
modified.lacZ      TGGCGGTTGGCCGAAAACTTGTCTGTCACCTTGCCAGCCGCCTCTCACGCCATCCCACAC
                   *****  **** ****** *   *** **  * ** ***** ** *********** **
```

```
original.lacZ      CTGACCACCAGCGAAATGGATTTTTGCATCGAGCTGGGTAATAAGCGTTGGCAATTTAAC
modified.lacZ      TTGACCACCTCTGAAATGGACTTCTGTATCGAATTGGGTAACAAGAGATGGCAATTCAAC
                    ********   ******** ** ** *****  ******* *** * ******** ***

original.lacZ      CGCCAGTCAGGCTTTCTTTCACAGATGTGGATTGGCGATAAAAAACAACTGCTGACGCCG
modified.lacZ      CGACAATCTGGCTTCCTTTCTCAAATGTGGATCGGCGACAAGAAGCAATTGTTGACCCCA
                   ** ** ** ***** ***** ** ******** ***** ** ** *** ** **** **

original.lacZ      CTGCGCGATCAGTTCACCCGTGCACCGCTGGATAACGACATTGGCGTAAGTGAAGCGACC
modified.lacZ      TTGAGGGACCAATTCACCAGAGCCCCATTGGACAACGACATCGGCGTCTCTGAAGCCACC
                    ** * ** ** ****** * ** **  **** ******** *****   ****** ***

original.lacZ      CGCATTGACCCTAACGCCTGGGTCGAACGCTGGAAGGCGGCGGGCCATTACCAGGCCGAA
modified.lacZ      AGGATCGACCCAAACGCCTGGGTCGAAAGGTGGAAGGCCGCCGGCCACTACCAAGCCGAA
                    * ** ***** *************** * ******** ** ***** ***** ******

original.lacZ      GCAGCGTTGTTGCAGTGCACGGCAGATACACTTGCTGATGCGGTGCTGATTACGACCGCT
modified.lacZ      GCCGCCTTGTTGCAATGTACCGCCGACACCCTTGCCGACGCCGTCTTGATCACCACCGCC
                   ** ** ******** ** ** ** ** ** ***** ** ** **  **** ** *****

original.lacZ      CACGCGTGGCAGCATCAGGGGAAAACCTTATTTATCAGCCGGAAAACCTACCGGATTGAT
modified.lacZ      CACGCCTGGCAACACCAAGGTAAGACCTTGTTCATCTCTAGAAAGACCTACAGGATCGAC
                   ***** ***** ** ** ** ** ***** ** ***    * ** ****** **** **

original.lacZ      GGTAGTGGTCAAATGGCGATTACCGTTGATGTTGAAGTGGCGAGCGATACACCGCATCCG
modified.lacZ      GGTTCTGGTCAAATGGCCATCACCGTCGACGTCGAAGTCGCCTCTGACACCCCACACCCA
                   ***  ************ ** ***** ** ** ***** **    ** ** ** ** **

original.lacZ      GCGCGGATTGGCCTGAACTGCCAGCTGGCGCAGGTAGCAGAGCGGGTAAACTGGCTCGGA
modified.lacZ      GCCAGGATCGGCTTGAACTGCCAATTGGCCCAAGTCGCCGAAAGGGTCAACTGGTTGGGT
                   **  **** *** **********  **** ** ** ** **  **** ****** * **

original.lacZ      TTAGGGCCGCAAGAAAACTATCCCGACCGCCTTACTGCCGCCTGTTTTGACCGCTGGGAT
modified.lacZ      TTGGGTCCACAAGAAAACTACCCAGACAGACTTACCGCCGCCTGTTTCGACAGATGGGAC
                   ** ** ** *********** ** *** * ***** *********** *** * *****

original.lacZ      CTGCCATTGTCAGACATGTATACCCCGTACGTCTTCCCGAGCGAAAACGGTCTGCGCTGC
modified.lacZ      TTGCCATTGTCTGACATGTACACCCCATACGTCTTCCCATCTGAAAACGGTTTGAGGTGC
                    ********** ******** ***** ***********    ********* ** * ***

original.lacZ      GGGACGCGCGAATTGAATTATGGCCCACACCAGTGGCGCGGCGACTTCCAGTTCAACATC
modified.lacZ      GGTACCAGAGAATTGAACTACGGCCCACACCAATGGCGGGGCGACTTCCAATTCAACATC
                   ** **  * ******** ** *********** ***** *********** *********

original.lacZ      AGCCGCTACAGTCAACAGCAACTGATGGAAACCAGCCATCGCCATCTGCTGCACGCGGAA
modified.lacZ      TCTCGCTACTCTCAACAACAATTGATGGAAACCTCTCACAGGCACTTGTTGCACGCCGAA
                      ******  ****** *** ***********   **  * **  ** ******* ***
```

```
original.lacZ      GAAGGCACATGGCTGAATATCGACGGTTTCCATATGGGGATTGGTGGCGACGACTCCTGG
modified.lacZ      GAAGGCACCTGGTTGAACATCGACGGTTTCCACATGGGTATCGGTGGCGACGACTCTTGG
                   ******** *** **** ************** ***** ** ************** ***


original.lacZ      AGCCCGTCAGTATCGGCGGAATTCCAGCTGAGCGCCGGTCGCTACCATTACCAGTTGGTC
modified.lacZ      TCTCCATCTGTATCTGCCGAATTCCAATTGAGCGCCGGTAGATACCACTACCAATTGGTC
                      ** ** ***** ** ********  *********** * ***** ***** ******


original.lacZ      TGGTGTCAAAAATAA
modified.lacZ      TGGTGCCAAAAGTGA
                   ***** ***** * *
```

4.5.9 Appendix II

Primers for amplification of E. coli tRNA genes – In addition to the shown sequences, each primer contains a 5′ end carrying the recognition sequence of *XmaI*.

tRNA Gene	Forward Primer	Reverse Primer
tyrU:4173495-4173579	CACCAGTTCGATTCCGGTAG	ACTTATCGTCTCGGGCTACG
tyrT:1286761-1286845	GGGAGCAGGCCAGTAAAAAG	TCTCACCGAAGTTACCACATC
thrU:4173411-4173486	TGAACTCGCATGTCTCCATAG	CTTTGGCCGCTCGGGAAC
thrT:4173777-4173852	GATGATGCGGGTTCGATTC	GAAGGAAAAAACAGGGAGGAG
serW:925107-925194	CCACCCATGAGGTTTGGTAG	AAAAAAAGCTCGCACTTTCG
serV:2816575-2816667	AGCACTCGTAAGAGGCGTGT	CGTAGCCGAGTACTCTATCCAG
serU:2041492-2041581	CAAATTTCCTGGCATCATGG	CGGGAAGTCGGGAGATAAG
proL:2284233-2284309	ATCGGTGTGGAAAACGGTAG	CCGTAAGGGTTGGTTTTTTC
proK:3706639-3706715	CGTATCTGCGCAGTAAGATGC	AAAAAAGCCTGCTCGTTGAG
leuX:4494428-4494512	ACAACGTTTTCCGCATACCT	CCTCAGTTGAGGTCTATTTACATACTTT
leuU:3320094-3320180	GACCAGCGATATCCCGAAC	TTTTCAGCGTCTCTTTTCTGG
leuP:4604223-4604309	CGTTGATATTGCTCGCACTG	CGCACAGTCATCTTACTTTTTTTG
ileX:3213620-3213695	GGATTGCGACACGGAGTTAC	GATTTCTCGTCAGCCTTTGC
ileT:4035164-4035240	AGATTGTCTGATGAAAATGAGCA	AACATGTAGTTAAAACCTCTTCAAA
glyV:4390383-4390458	GAAATGCGAAAATTACGAAAGC	GGTGGTCTGTGCTTTGCAG
glyU:2997006-2997079	AAGGAGAGCGTAAGGTTTATAATG	GGGGAAGTATTACGGCGAAG
glyT:4173696-4173770	AAAATCAGGTAGCCGAGTTCC	AAGGGTGCGCTCTACCAAC
glnV:695765-695839	TGTTCGGCAAATTCAAAACC	CAACTGGGTGCACTTACAAGG
glnU:696088-696162	CGCACCATTCACCAGAAAG	AATAACCGGGCGGTGAAC
argX:3980398-3980474	TGGGAAGTCCGTATTATCCAC	TACTACCACCGCAGCTCAAG
argW:2464331-2464405	TACCCCGCACTCCATTAGC	ATTTTGCGGACTGGTACGG
argU:563946-564022_4	ACGCGATTACACCGCATTG	TGGAGGATATAAAGAAGGCGTAAC
argQ:2815806-2815882	AGCGGTATCAATATCAGCAGTAG	TCTCTTCGATACCTTCATTGC
alaW:2516178-2516253	CCCGTCAACTCGACAAGC	CGATGCGTTTACGTACCAAG
alaT:4035283-4035358_1	GGCAAATTTGAAGAGGTTTTAACTAC	GGTACACTCTGAAGTATTTTTTATTTAATC

4.5.10 Appendix III

The sequence of the inserted fragment carried by pBAD-KAN (the plasmid that was selected in the presence of kanamycin) – Ten clones were analyzed for their sequences, all of which contained one copy of *serW* gene in reverse direction (blue) and one copy of *glyT* in forward direction (red). *XmaI* digestion sites are underlined.

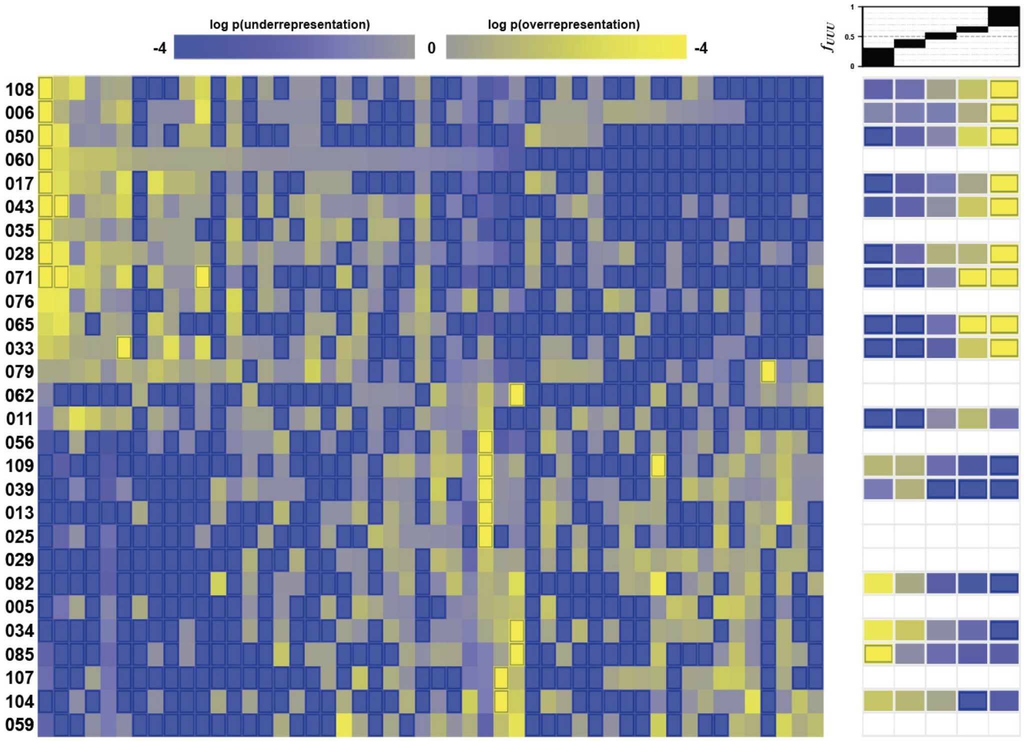
TGACNCTGAGCTTTTATCGCAACTCTCTACTGTTTCTCCATACCCGTTTTTTTGGGCTAGC
GAATTCGAGCTCGGTACCCGGGAAAAAAGCTCGCACTTTCGTACGAGCTCTTCTTTAAAT
A**TGGCGGTGAGGGGGGGATTCGAACCCCCGATACGTTGCCGTATACACACTTTCCAGGCGT**
GCTCCTTCAGCCACTCGGACACCTCACCAAATTGTTTTGCTACCAAACCTCATGGGTGGCC
CGGGAAAATCAGGTAGCCGAGTTCCAGGAT**GCGGGCATCGTATAATGGCTATTACCTCAGC**
CTTCCAAGCTGATGATGCGGGTTCGATTCCCGCTGCCCGCTCCAAGATGTGCTGATATAGC
TCAGTTGGTAGAGCGCACCCTTCCCGGGGATCCTCTAGAGTCGACCTGCAGGCATGCAAGC

4.5.11 Appendix IV

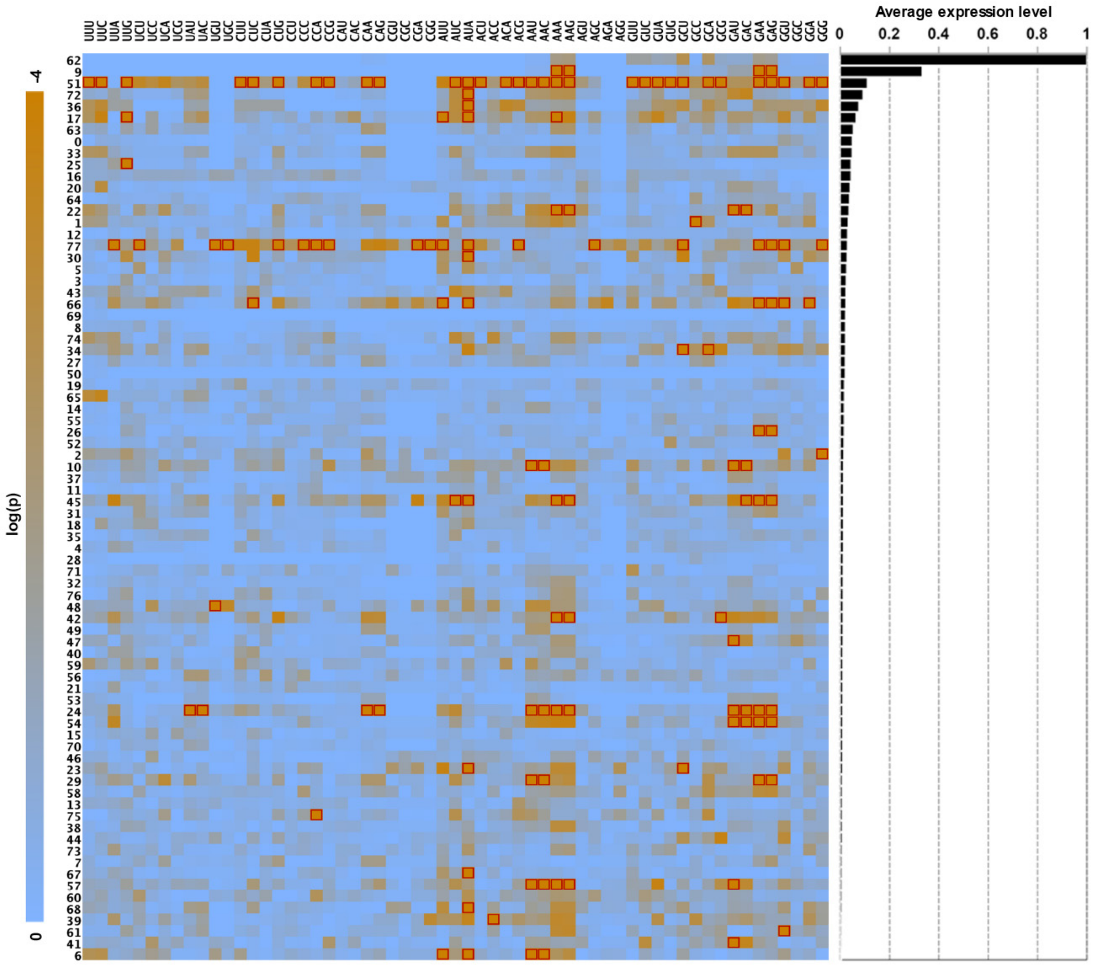
The sequence of the inserted fragment carried by pBAD-TET (the plasmid that was selected in the presence of tetracycline) – Ten clones were analyzed for their sequences, all of which contained one copy of *ileX* gene in forward direction (red). *XmaI* digestion sites are underlined.

TCGCACTCTCTACTGTTTCTCCATACCCGTTTTTTTGGGCTAGCGAATTCNAGCTCGGTAC
CCGGGGATTGCGACACGGAGTTACTTTATAATCCGCTACCAT**GGCCCCTTAGCTCAGTGGT**
TAGAGCAGGCGACTCATAATCGCTTGGTCGCTGGTTCAAGTCCAGCAGGGGCCACCAGATA
TAGCAAAGGCTGACGAGAAATCCCCGGGGATCCTCTAGAGTCGACCTGCAGGCATGCAAGC

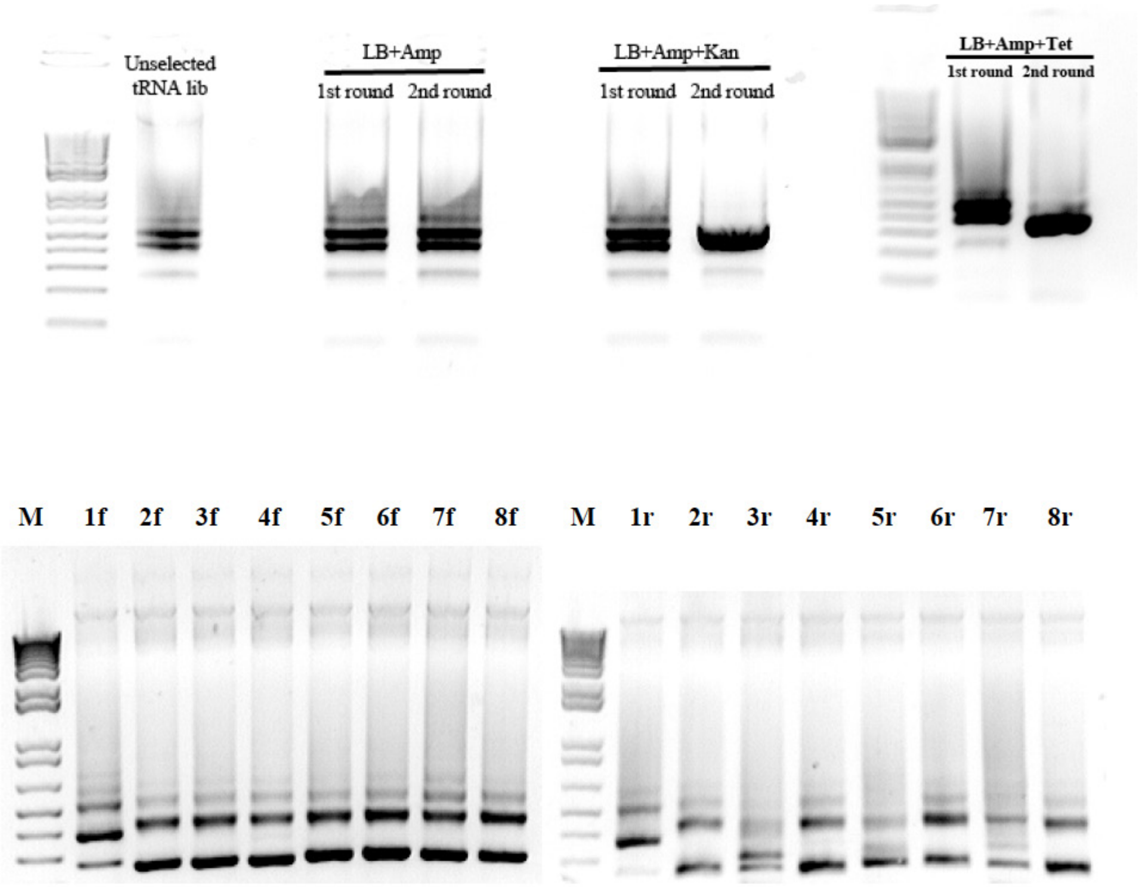
4.6 Supplementary Figures



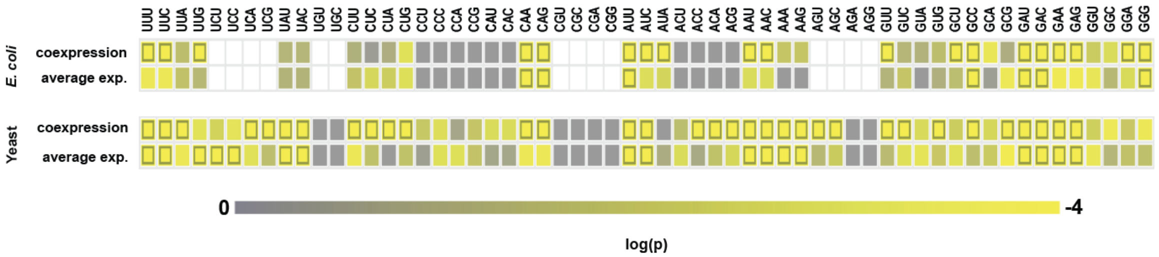
Supplementary Figure 4-1. clustering genes based on their synonymous codon usage results in specific enrichment of different expression profiles – (Left panel) Human genes were clustered based on their synonymous codon usage, exploiting a modified k-means algorithm that could handle missing data. The mutual information (MI) of distribution of each expression profile among different codon usage clusters was examined. Expression profiles that showed significantly non-random distributions based on their MI values ($p \leq 1 \times 10^{-4}$) are shown in this figure. Each row represents one coexpression cluster, whose number, according to [168], is shown at left. Each column stands for one of the codon usage-based clusters. **(Right panel)** codon UUU is over-represented in some coexpression clusters and under-represented in others. Only clusters whose mutual information for UUU is significantly higher than random ($p \leq 1 \times 10^{-4}$) are shown. In both panels, a yellow frame around a square means significant over-representation of the corresponding expression ($p \leq 1 \times 10^{-4}$), and a blue frame around a square means significant under-representation ($p \leq 1 \times 10^{-4}$).



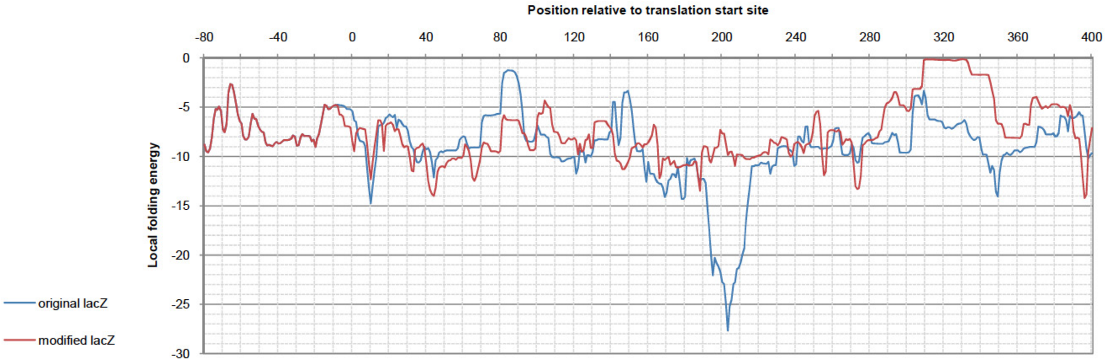
Supplementary Figure 4-2. Mutual information of synonymous codon usage in yeast coexpression – Each row represents a coexpression cluster with the ID number of the cluster written on left, while each column represents a codon. Stop codons, AUG and UGG are omitted. Significant mutual information (MI) values ($p \leq 1 \times 10^{-4}$) are highlighted by red frames. The average expression level of genes within each coexpression cluster is shown on the right chart, normalized to the maximum of 1.0. Coexpression cluster ID numbers are the same as reported in [168].



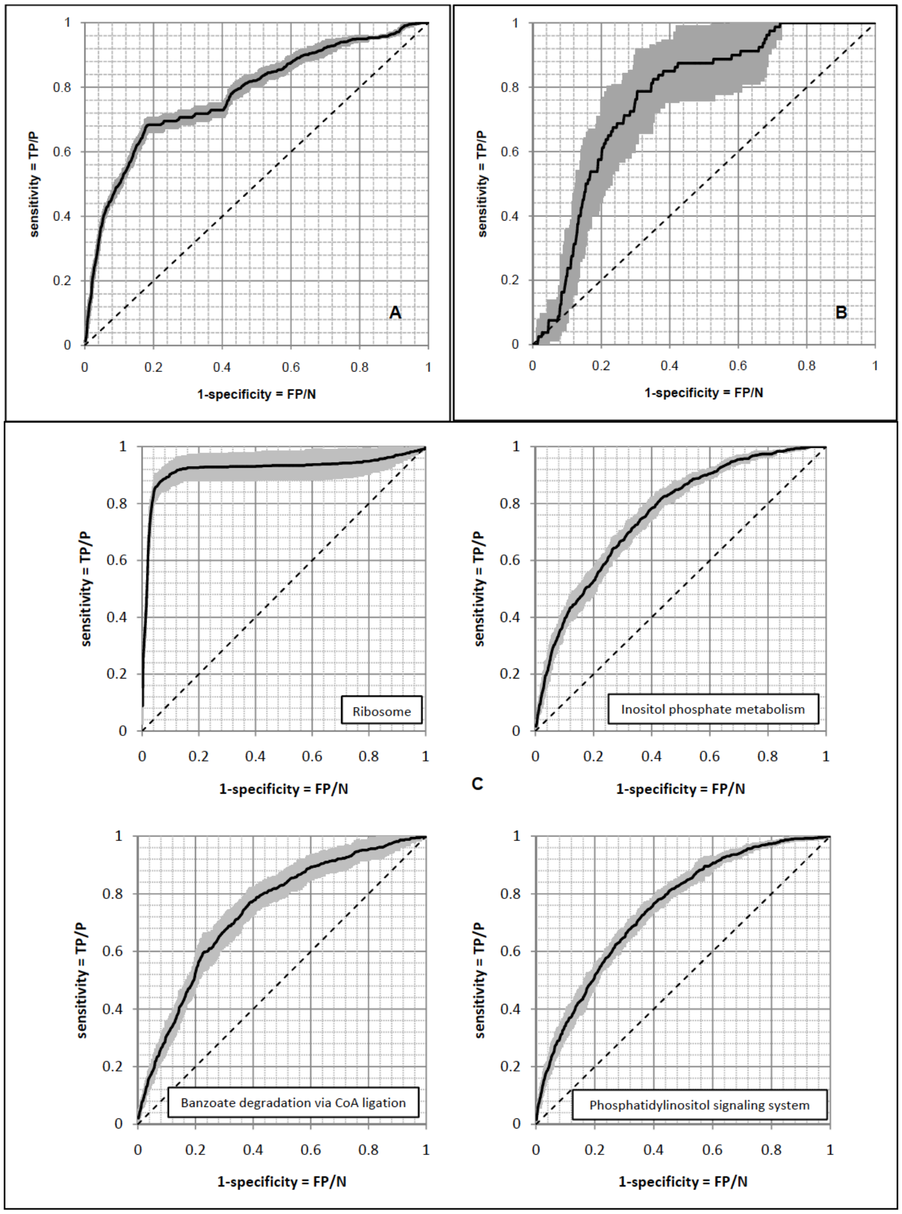
Supplementary Figure 4-3. Stress conditions can select a particular plasmid from the tRNA library – (Top panel) Unselected tRNA library results in a ladder-like pattern on the gel after cloning site amplification, with amplicons containing no tRNA genes at the bottom and amplicons containing several tRNA genes at the top. Culturing the library in LB+Amp medium (negative control) does not change this pattern (pBAD18 has an Amp^r marker; thus, presence of Amp is not considered a selection pressure). This is while after two passages in LB+Amp+Kan, a single plasmid is enriched in the library. A single plasmid is also selected in the presence of tetracycline (LB+Amp+Tet). **(Bottom panel)** The tRNA library represents different tRNA isoacceptors of Gly, Ile and Ser in both forward and reverse directions. Each PCR amplification results in more than one fragment since a tRNA gene may be placed in different combinations with respect to other tRNA genes: directly after annealing site of the first primer, or isolated from this site by one or more other tRNAs. 1-8: *glyV*, *glyU*, *glyT*, *ileX*, *ileT*, *serW*, *serV* and *serU*, respectively; f: forward; r: reverse; M: standard 1kb plus ladder (Invitrogen).



Supplementary Figure 4-4. The significance of correlation coefficient between codon usage and either average expression level or expression pattern, after correcting for confounding effect of each factor – The correlation coefficient between codon usage and each factor was evaluated in a multivariate analysis, including the other factor as the covariate. Thus, for example, in the first row a red frame indicates that the correlation between codon usage and expression pattern in *E. coli* is significant even after the confounding effect of average expression level is considered. Although it is apparent that average expression level correlates with expression pattern in both *E. coli* and yeast, there are many cases in which correlation of expression pattern with codon usage only partly overlaps with correlation of average expression level and codon usage. It is notable that in yeast, 33 out of 59 codons have significant correlation with expression pattern after correcting for the confounding effect of average expression level, while only 17 codons have significant correlation with average expression level after removing the confounding effect of expression pattern. This indicates that expression pattern can more strongly explain the synonymous codon usage.



Supplementary Figure 4-5. Local folding energy of original and modified versions of *lacZ* mRNA – Folding energy was calculated for sliding windows of 40nt, including the 5′ UTR of each mRNA (the results are shown only for the 5′ UTR and the first 400nt of the coding sequence). The folding energies of the two variants are especially similar in the first 40nt of the coding region; this region has been found recently to have the strongest correlation with expression level [187].



Supplementary Figure 4-6. Prediction of expression profile and function in human and a *Trypanosoma brucei*, based on synonymous codon usage – These examples illustrate that genes may be successfully assigned to coexpression clusters and/or functions, solely based on their synonymous codon usage. This indicates that codon usages of different coexpression/function clusters are different, allowing their discrimination from each other. Panel (A) belongs to coexpression cluster 50 from human [168] which mostly consists of proteins involved in mRNA metabolic process (GO:0016071). Panel (B) belongs to base pair excision repair proteins (KEGG:hsa03410). Panel (C) corresponds to homology-independent prediction of gene function in *Trypanosoma brucei*. *T. brucei* is the causative agent of human African trypanosomiasis. Almost 55% of the genes in the genome of *T. brucei* do not have any homologs outside of trypanosomatid clade. Thus, their functions cannot be determined by homology-dependent methods. Out of seven functional clusters that had at least 40 *T. brucei* genes in KEGG pathway, four could be predict by our naïve Bayesian network with reasonable accuracy. In each panel, the standard deviation of sensitivity is illustrated by the shaded region. In all cases, a two-fold cross-validation was used, where each time half of the genes were selected randomly and used for training the classifier, while the other half was reserved for assessing the prediction power. Standard deviation of sensitivity, indicated as a gray region in each panel, was calculated by repeating the cross-validation 10 times. The diagonal lines represent the expected performance if codon usage was not able to classify the genes.

5 Genome-wide computational identification of functional RNA elements in *Trypanosoma brucei* and their application in gene function prediction

The previous chapter introduced codon usage as an expression regulator, which correlates with gene function in a wide range of organisms. As shown in Supplementary Figure 4-6, this function-specificity can be utilized to predict gene function. However, the performance of this prediction method is not satisfactory on its own, and needs to be improved by integrating with orthogonal sources of information. In the next chapter, which was published as an article in BMC Genomics in 2009 [3], we describe prediction of functional RNAs, including non-coding RNAs (ncRNAs) and *cis*-acting RNA elements involved in post-transcriptional gene regulation, based on two independent computational analyses of the genome of *Trypanosoma brucei*. We then discuss the utility of these regulatory elements in homology-independent function prediction in *T. brucei*. This first genome-wide analysis of fRNAs in trypanosomatids restricts the search space of experimental approaches and, thus, can significantly expedite the process of characterization of functional RNA elements. Our classifiers for function prediction based on *cis*-acting regulatory elements can also, in combination with other methods, provide the means for homology-independent annotation of trypanosomatid genomes.

5.1.1 Background

RNA elements that are functional at RNA level, i.e., functional RNAs (fRNAs), are becoming to be appreciated more and more as their diverse structural, regulatory and catalytic roles are revealed [199, 200]. Several classes of fRNAs have been identified, including different types of non-coding RNAs (ncRNAs) such as tRNAs, rRNAs, microRNAs (miRNAs), telomerase RNA, RPR1 (the RNA component of nuclear RNase P), small nuclear RNAs (snRNAs) and small nucleolar RNAs (snoRNAs). The *cis*-regulatory elements in the 5′ - and 3′-untranslated regions (UTRs) of mRNAs constitute another class of fRNAs that are mostly involved in post-transcriptional regulation of gene expression (see [31, 32]). Recent developments in computational tools for prediction of fRNAs have shown a widespread set of RNA elements that are specifically involved in post-transcriptional regulatory processes [201]. Although crucial in many different species, post-transcriptional regulation is especially the major mechanism for regulation of gene expression in a group of unicellular parasites called trypanosomatids.

Trypanosomatids, including *Trypanosoma brucei*, *T. cruzi* and different *Leishmania* species, are the causative agents of serious human as well as animal diseases, with a very high incidence and mortality rate if untreated. Genes in trypanosomatids are transcribed as polycistronic mRNAs [93] that are further processed via trans-splicing [47]. Regulation of gene expression, which occurs mostly during or after splicing, involves several *cis*-acting fRNA elements, such as U-rich elements (UREs), short interspersed degenerated retroposons (SIDERs), etc. [31, 32]. These elements mostly regulate either the stability or translation rate of mRNAs via interaction with different *trans*-acting proteins, many of which are unknown. It has also been proposed recently that miRNAs may play a role in posttranscriptional gene regulation in *T. brucei* [202], although no experimental substantiation has been found.

Experimental identification of *cis*-acting fRNA elements is an exhausting task that requires extensive functional assays with several strains carrying deletion/substitution mutants of a likely regulatory sequence. The situation is not better for ncRNAs, as it is not clear in which region(s) in the genome they should be searched for and for what particular function the screening experiment should be designed (as opposed to *cis*-acting fRNA elements that occur adjacent to coding sequences and affect gene expression). Although computational identification of fRNAs from genome sequences can be an alternative, it is not yet as robust as identification of protein-coding RNAs, due to the lack of strong conserved signals in their sequences [203]. Here, we present a computational examination of the genomes of *T. brucei* and *L. braziliensis* in order to identify a set of conserved ncRNAs that, based on computational and statistical analysis, are highly reliable. We show that our methodology is able to find a large number of known as well as novel potential ncRNAs. We further examine our candidate ncRNAs for the presence of potential pre-miRNAs, and show that the existence of miRNA genes that are conserved between *T. brucei* and *L. braziliensis* is highly unlikely. We also use a different method for homology-independent identification of short regulatory RNA motifs in 5′ and 3′ UTRs of *T. brucei* genes. These motifs complement our predicted ncRNAs by providing a set of the most functionally important regions of potential *cis*-regulatory fRNA elements. In addition to offering new insights about the regulatory mechanisms of protein expression in *T. brucei*, these regulatory motifs can be used for prediction of gene function.

5.2 Results and Discussion

5.2.1 Identification of conserved ncRNAs in T. brucei

We compared the genome sequences of *T. brucei* and *L. braziliensis* in order to identify conserved genomic regions. *L. braziliensis* is the only trypanosomatid other than *T. brucei* with available genome sequence in which the putative components of RNAi machinery have been identified [27]. Thus, its comparison with *T. brucei* provides the possibility of detecting conserved ncRNAs involved in or processed by this machinery. We used a binomial-based model [204] to assess the conservation across *T. brucei* genome in comparison to the genome sequence of *L. braziliensis*. Using this model, we found that about 18% of the *T. brucei* genome shows conservation degrees above the median that would be expected from a random distribution. These regions, in addition to being enriched for functional elements, have allegedly the highest-quality alignments compared to the alignments that correspond to less conserved regions. This conserved subset of the *T. brucei* genome consisted of about 5.26 Mbp of protein-coding sequences and 887 kbp of non-coding sequences. We used QRNA [205] to identify parts of these conserved genomic regions that showed patterns of conserved structural RNA elements. About 37.2 kbp of the non-coding conserved genomic regions obtained RNA scores above zero, using QRNA. For the protein-coding conserved regions, this number was about 16.8 kbp, indicating a false positive rate of about 0.3%. Assuming this false positive rate, we would expect about 2.8 kbp of false positives among non-coding genomic regions and, hence, a precision of about 92.3% (precision was defined as TP/(TP+FP), where TP and FP stand for the number of true positives and false positives among non-coding genomic regions, respectively).

It should be noted that the estimated false positive rate from coding sequences would not be applicable to non-coding sequences if we had included scores other than the RNA score from QRNA, such as the COD and OTH scores (COD and OTH scores express the likelihood of being a coding sequence and a non-RNA, non-coding sequence, respectively). However, the behavior of QRNA may still be different between coding regions and non-RNA, non-coding genomic regions as coding sequence evolves in a very specific way. Furthermore, RNA structure in coding sequence may be specifically selected against. We have thus used a different, more conservative method for estimating the false positive rate of our ncRNA predictions, which is explained in the section "Identification of highly significant candidate ncRNAs"

About 5.2 kbp of our found candidate fRNAs overlapped with already annotated rRNA, snRNA, and tRNA genes, indicating the capability of our approach in finding non-coding RNAs. The sensitivity of this approach, i.e., TP/(TP+FN) where FN indicates the number of false negatives, showed considerable differences among different classes of structural RNAs. For example, 30 of our predicated candidates overlapped one of the 65 known tRNAs, equal to about 50% sensitivity for detection of tRNAs. On the other hand, only 21 candidates overlapped one of the 106 known rRNA genes, indicating a lower sensitivity for rRNA detection. This is while we detected none of the 353 known small nucleolar RNAs (snoRNAs). This may indicate the lack of conservation of snoRNA structure between *T. brucei* and *L. braziliensis*.

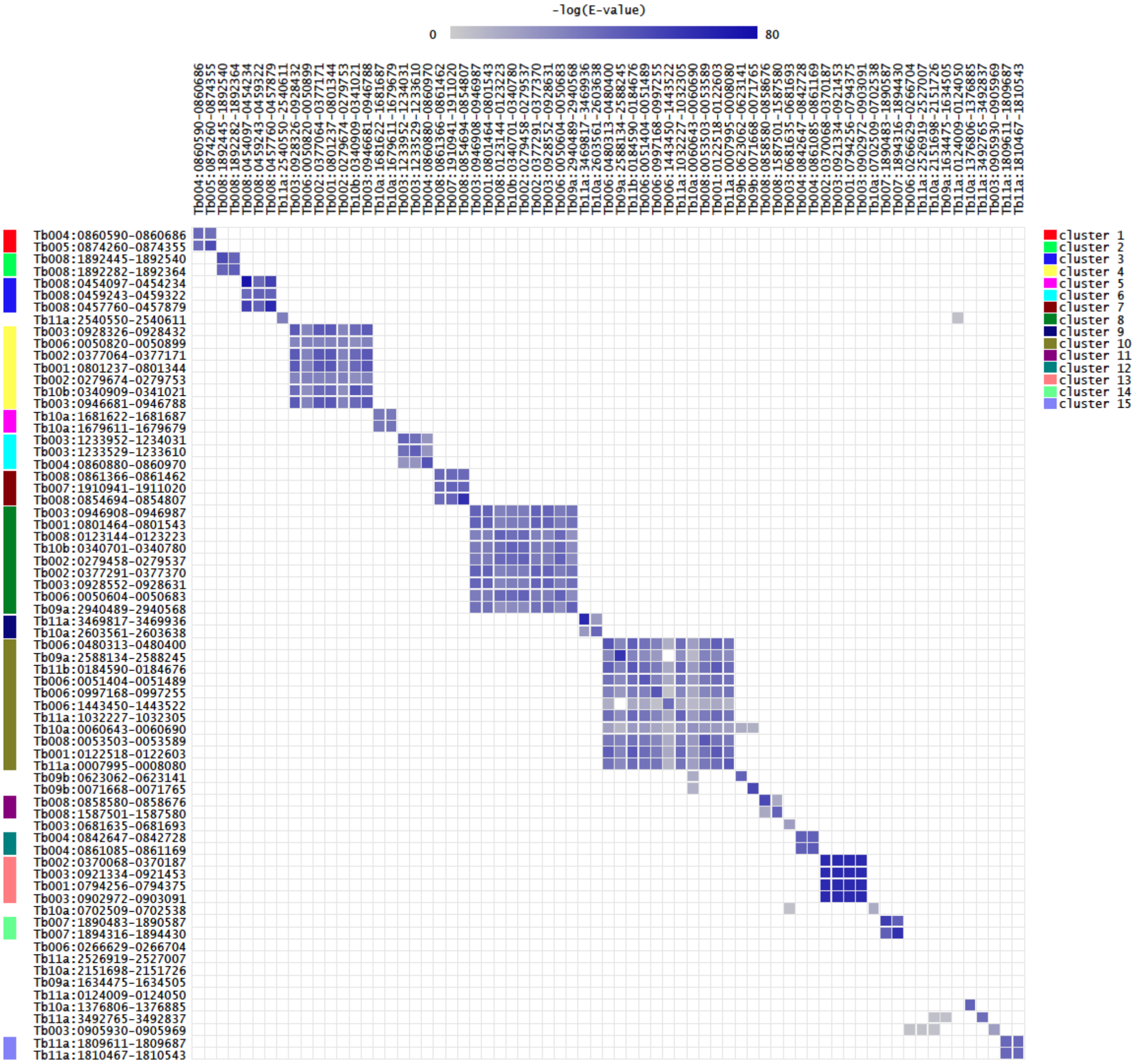


Figure 5-1. Homology table for the predicted ncRNAs – Many candidate ncRNAs can be grouped into several homology clusters, here shown by color labels (clusters 1-15). In this figure, only ncRNAs are shown for which there is at least one other predicted ncRNA with homology E-value < 0.0025 and alignment coverage > 50%. The color of each square reflects the BLAST E-value with the sequence in the corresponding row as the query.

A complete list of all found ncRNA candidates along with their associated information can be found in the online Additional File 1 at http://www.biomedcentral.com/1471-2164/10/355/additional/. Many of these candidates can be grouped into several homology clusters, as shown in Figure 5-1. When several homologous sequences are independently predicted to be ncRNAs, the predictions can be considered highly reliable. Sequences within clusters 1, 2, 6, 7, 9, 11, 12 and 14 either overlap with or are homologous to known tRNAs. Similarly, sequences within clusters 3, 4, 8 and 13 seem to represent rRNAs. However, clusters 5, 10 and 15 do not correspond to any known ncRNAs and, thus, may represent novel ncRNA classes with unknown functions. Cluster 10 is of particular significance due to its large size, indicating that the elements of this class may be present at a high frequency in the genome.

5.2.2 Investigating the presence of conserved miRNA genes

Based on a computational analysis of *T. brucei* genome, it has been recently proposed that trypanosomatids may use miRNAs in order to regulate the levels of particular mRNAs [202]. However, this report is not consistent with our current knowledge of miRNA origin [206, 207]; regulation via miRNA seems to have emerged in a completely different branch of life, although its convergent evolution in several branches is not impossible. Hence, we decided to investigate the presence of putative miRNA precursors among our predicted ncRNAs through a relatively simple, yet specific approach that considers a few structural and thermodynamic criteria for identification of pre-miRNA sequences (see Methods section). Using 250 pre-miRNAs that, as control sequences, were randomly selected from 24 different organisms (Supplementary Table 5-1), it can be estimated that the sensitivity of pre-miRNA prediction using our criteria is about 32.4% ± 2.1%. Also, using a set of 30770 randomly selected sequences from *T. brucei* genome, the specificity of this method can be estimated at about 99.1% ± 0.3%.

After removing low-complexity regions (LCRs, see the Methods section), only five of the predicted ncRNA sequences met our criteria for structure and free folding energy (Supplementary Figure 5-1). However, these rare sequences mostly consist of dinucleotide repeats (particularly AU repeats), and can be accounted for false positives of our method. Based on our analysis, it is highly unlikely to expect any conserved miRNA genes in *T. brucei*. It should be mentioned that a large number of the previously predicted *T. brucei* miRNAs [202] were potentially targeting variant surface glycoproteins (VSGs), which are absent in *L. braziliensis*. However, other predicted miRNAs were targeting conserved complexes such as 20S proteasome, and thus would be expected to be found in this study if they were conserved. Although this analysis does not definitely reject the presence of miRNAs in *T. brucei* genome, suggests that a reexamination of this genome for the presence of such elements is required.

5.2.3 Identification of highly significant candidate ncRNAs

In order to select a highly significant subset from our set of candidate conserved ncRNAs, we filtered out the candidates whose QRNA scores were not significantly higher than expected from a random distribution. The random distribution for each candidate ncRNA was obtained by computing the QRNA scores of 1000 randomly scrambled *T. brucei-L. braziliensis* alignments, as described in the Methods section. A candidate ncRNA was rejected if it was outscored by more than three randomized versions (i.e., $p \leq 0.003$; this *p*-value threshold was selected so as the expected number of false positives would be less than one). This filtering procedure resulted in 117 highly significant novel putative ncRNAs (online Additional File 4 and the first 117 candidates in online Additional File 1 at http://www.biomedcentral.com/1471-2164/10/355/additional/), of which 53 neither overlapped nor were homologous to any annotated features of *T. brucei* genome and, hence, may represent completely novel ncRNAs (Table 5-1). All 117 candidates that did not overlap with a coding sequence had the highest score for the RNA model and not the COD and OTH models, although they were initially selected only based on their RNA scores and irrespective of their COD and OTH scores.

The calculated *p*-value provides another measure, though more conservative, for estimating the precision of our method. For example, a *p*-value ≤0.001 is equal to about 0.887 kbp of false positives (out of 887 kbp of the non-coding conserved genomic regions), assuming that most of the non-coding genome consists of non-RNA random sequences. This is while more than 5.7 kbp of our candidates (the top 79 candidates in online Additional File 1 at http://www.biomedcentral.com/1471-2164/10/355/additional/) were significant at this level, indicating a precision of about 85%.

Table 5-1. Classification of predicted ncRNAs in *T. brucei* genome – Candidate ncRNAs are classified based on either homology with known ncRNAs or overlap with known genomic features. Candidate ncRNAs within each class are further divided into subgroups based on their location relative to known genomic features.

Classification	Within 100 bp of a non-overlapping coding sequence**	Within/flanking ncRNA cluster (no. of ncRNAs >2)***	Within a strand switch region***	Elsewhere	**Total**
Overlap CDS	0	0	0	36	**36**
Overlap pseudogene	0	0	0	1	**1**
Overlap unlikely proteins	0	0	0	0	**0**
Homologous to rRNA*,**	0	4	2	1	**5**
Homologous to tRNA*	1	0	0	0	**1**
Overlap known ncRNA**	0	25	7	1	**26**
Overlap Ingi/RIME repeat	0	0	0	0	**0**
Unclassified	8	1	1	43	**53**
Total	**9**	**26**	**8**	**81**	**117**

* Each candidate may contain several closely located single ncRNAs, some of which may have already been annotated on the current release of the *T. brucei* genome. However, at least one ncRNA within each sequence is unannotated, for which a known homolog is found. These unannotated ncRNAs represent novel instances of their classes.

** These categories may overlap.

*** These categories may overlap.

These novel ncRNAs did not show any statistically significant enrichment in particular genomic positions such as regions with clustered ncRNAs, strand switch regions (regions where the coding strand changes) or regions adjacent to coding sequences (significance was defined as p-value < 0.05 in a genomic position permutation test), indicating a relatively uniform distribution on the genome. Nonetheless, an educated guess can be made for biological functions of some candidates based on their positions. For example, eight unclassified candidate elements occur in the vicinity of a coding sequence. These elements may represent regulatory structures at 5′ or 3′ UTRs of coding sequences, involved in post-transcriptional regulation of gene expression. Also, one unclassified candidate fRNA was found to occur in a strand switch region. As transcription of polycistronic mRNAs start from strand switch regions, this fRNA may represent an element in the 5′ end of the resultant transcript, and may be involved in its localization, posttranscriptional processing or regulation.

Expectedly, none of the previously characterized *cis*-regulatory RNA elements of *T. brucei* were found among our set of candidate structural RNA elements. This is not surprising since the known regulatory RNA elements of *T. brucei* are not conserved in *Leishmania* species [31, 32]. Furthermore, many of these elements are known via their sequence, not their structure. We specifically discuss the computational identification of *cis*-regulatory RNA elements in *T. brucei* in the next section.

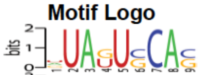
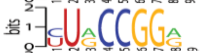
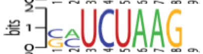
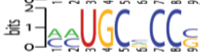
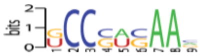
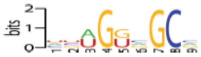
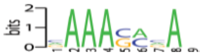
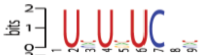
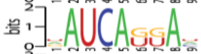
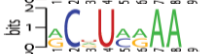
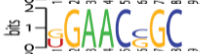
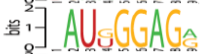
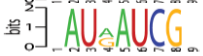
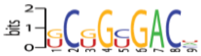
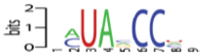
5.2.4 Finding informative function-specific regulatory elements

We used a homology-independent approach to investigate the presence of function-specific motifs in 5 and 3′ UTRs of *T. brucei* genes, using a recently developed algorithm, named FIRE [168]. It has been shown that FIRE is able to identify many known and novel regulatory elements, with a near-zero false positive discovery rate, in upstream and downstream of genes that are clustered according to their expression patterns. Here, we used FIRE to find ‘function-specific’ regulatory elements in 5and 3′ UTRs of *T. brucei* genes: genes with similar functions are usually co-regulated [74], indicating that they should have similar *cis*-regulatory elements. Thus, clustering genes according to their functions can be used as a surrogate of clustering them according to their expression patterns. This approach is particularly useful for organisms in which gene regulation occurs mostly at post-transcriptional levels, such as trypanosomatids (transcript profiling studies have been suggested to be unable to identify the dynamics of protein expression in such organisms; see [208]). We were able to identify 15 function-specific motifs in 5′ UTRs of *T. brucei* genes and 21 function-specific motifs in their 3′UTRs (Table 5-2, Table 5-3, and Supplementary Figure 5-2). Based on the results of running FIRE on 10 permuted sets of gene-function assignments, we can estimate an expected precision of 75.3% for discovering function-specific 5′ UTR motifs and 84.8% for 3′ UTR motifs.

Most of the motifs that are found by FIRE have orientation bias, i.e., mostly occur at a particular orientation with respect to the coding sequence. This property is expected from RNA motifs. Furthermore, two of the motifs that were predicted in 3′UTRs have position bias, which means that they prefer to be at a particular distance from the stop codon of the upstream coding sequence. This property has also been observed for many regulatory motifs in different organisms [74, 168], and further increases the possibility that the predicted motif has a biological role.

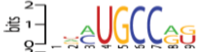
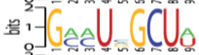
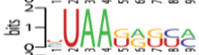
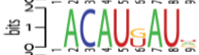
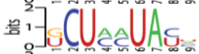
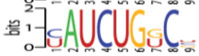
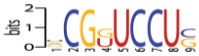
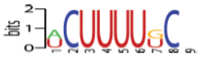
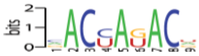
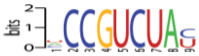
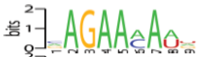
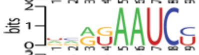
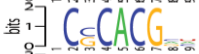
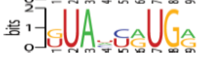
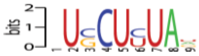
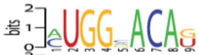
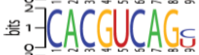
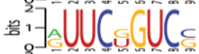
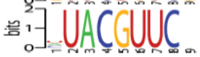
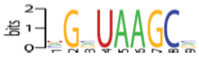
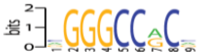
Our predicted function-specific motifs overlap with a number of experimentally found regulatory sequences in *T. brucei*, mostly identified by deleting different parts of UTRs and evaluating the effects of these deletions on regulation of a reporter gene: It has been shown that the 3′ UTR of glycosomal phosphoglycerate kinase PGKC can cause bloodstream form-specific gene expression [68, 209]. We found that this regulatory sequence contained six of our predicted 3′UTR motifs ($p<1\times10^{-5}$), most notably the glycolysis-specific motif VGGGCCRCV (degenerate positions are shown using IUPAC nomenclature of mixed bases [210]). Interestingly, the 5′ UTR of the same gene, which has been shown to affect splicing in procyclic stage [211, 212], also contains two copies of the 5′ UTR motif UHUDUCNH. As another example, the 3′UTR of fructose bisphosphate aldolase contains an instance of the fructose metabolism-specific motif MUGGVACAK. This untranslated region has also been reported to be able to cause regulated expression of genes in *T. brucei* [209].

Table 5-2. Function-specific motifs in 5′ UTRs of *T. brucei* genes – The functions in which each motif is significantly overrepresented or underrepresented are indicated in the second column using black and blue text colors, respectively.

Motif Logo	KEGG pathway	MI[a]	Z-score[b]	Robustness[c]	Position bias[d]	Orientation bias[e]
	Pyrimidine metabolism	0.026	26.403	5/10	-	-
	Pyrimidine metabolism	0.0293	31.957	4/10	-	→
	Oxidative phosphorylation	0.0297	34.309	7/10	-	-
	Ubiquitin mediated proteolysis	0.0277	29.528	6/10	-	←
	Inositol phosphate metabolism	0.0258	15.789	4/10	-	←
	Benzoate degradation via CoA ligation	0.0325	20.148	8/10	-	←
	Phosphatidylinositol signaling system	0.0254	15.536	3/10	-	←
	Inositol phosphate metabolism	0.0374	23.457	9/10	-	→
	Benzoate degradation via CoA ligation	0.0277	16.623	8/10	-	-
	Phosphatidylinositol signaling system	0.0409	25.678	10/10	-	→
	Ribosome	0.0254	15.979	6/10	-	→
	Proteasome	0.0415	26.471	9/10	-	→
	Ribosome	0.0655	43.401	10/10	-	-
	Aminoacyl-tRNA biosynthesis	0.0307	34.474	5/10	-	→
	Glycine, serine and threonine metabolism	0.0363	29.553	7/10	-	→
	Fructose and mannose metabolism	0.0334	33.335	7/10	-	→
	Alanine and aspartate metabolism	0.0392	42.428	7/10	-	→
	Carbon fixation	0.0316	29.288	8/10	-	←
	Propanoate metabolism	0.0477	49.555	9/10	-	-
	Limonene and pinene degradation	0.0207	19.954	3/10	-	-
	Glycolysis / Gluconeogenesis	0.0136	7.577	0/10	-	-
	Pyruvate metabolism	0.0389	22.893	6/10	-	→

a. Mutual information value **b.** Z-score associated with the MI value **c.** Robustness, obtained from ten jack-knife trials of randomly removing one-third of the genes and reassessing the statistical significance of the resulting MI value **d.** Position bias indicator ("Y" if a position bias is observed) **e.** Orientation bias, indicating the orientation of the motif with respect to its associated coding sequence

Table 5-3. Function-specific motifs in 3' UTRs – See **Table 5-2** for column explanations.

Motif Logo	KEGG pathway	MI	Z-score	Robustness	Position bias	Orientation bias
	Glycerophospholipid metabolism	0.0285	15.782	4/10	-	→
	Purine metabolism	0.0327	18.225	8/10	-	→
	Purine metabolism	0.0538	33.373	9/10	-	→
	Glycerolipid metabolism	0.0317	24.18	5/10	-	→
	SNARE interactions in vesicular transport	0.0285	15.918	7/10	-	→
	Ubiquitin mediated proteolysis	0.0212	22.356	4/10	-	→
	Inositol phosphate metabolism	0.0199	11.472	1/10	-	→
	Benzoate degradation via CoA ligation	0.0274	16.34	5/10	-	→
	Phosphatidylinositol signaling system	0.0246	14.37	4/10	-	→
	Inositol phosphate metabolism	0.0396	25.365	9/10	Y	→
	Benzoate degradation via CoA ligation	0.0347	21.319	7/10	Y	→
	Phosphatidylinositol signaling system	0.0415	26.458	9/10	Y	→
	Inositol phosphate metabolism	0.0202	11.061	1/10	-	→
	Phosphatidylinositol signaling system	0.0239	13.24	3/10	-	→
	Inositol phosphate metabolism	0.0315	29.029	5/10	-	→
	Benzoate degradation via CoA ligation	0.0211	20.469	2/10	-	→
	Ribosome	0.0535	33.667	10/10	-	→
	Aminoacyl-tRNA biosynthesis	0.0301	16.979	3/10	-	→
	Methionine metabolism	0.0255	16.214	6/10	-	→
	Aminosugars metabolism	0.0392	22.565	5/10	-	→
	Glycine, serine and threonine metabolism	0.0288	15.751	5/10	-	→
	Ribosome	0.0255	15.04	4/10	-	→
	Fructose and mannose metabolism	0.0269	29.239	2/10	-	→
	Fructose and mannose metabolism	0.0336	31.962	6/10	-	→
	Valine, leucine and isoleucine degradation	0.0303	32.565	5/10	-	→
	Butanoate metabolism	0.0205	21.474	3/10	Y	→
	Propanoate metabolism	0.0185	17.977	4/10	-	→
	Limonene and pinene degradation	0.0257	25.184	4/10	-	→
	Glycolysis / Gluconeogenesis	0.0311	22.104	6/10	-	→

It should be noted that our approach is only able to identify function-specific short RNA motifs, not motifs that are involved in regulation of expression in a rather genome-wide scope, or in a gene-specific manner. Thus, it is not surprising to see that some of the previously identified regulatory elements, such as the widely used U-rich elements [32] are not among our motifs. Structural RNA elements also cannot be identified using FIRE; nonetheless, some of our found short motifs may represent the most functionally important regions of RNA structural elements.

5.2.5 Function prediction using regulatory RNA motifs

We devised a naïve Bayesian network that based on the pattern of presence and absence of motifs in 5′ UTRs and 3′ UTRs can predict whether a gene belongs to a particular pathway (see the Methods section). For many pathways, this naïve Bayesian network can be used to classify *T. brucei* genes with acceptable reliability (see Figure 5-2 for an example). As it is shown in Figure 5-2A, only a few motifs are needed to reach the maximum possible prediction power. However, adding more motifs to this classifier does not reduce the prediction power, which simplifies the design of effective naïve Bayesian networks. We expect that by combining this method with other function prediction methods, we will be able to expand the functional annotations of *T. brucei* genes extensively. A complete assessment of function prediction in *T. brucei* using our method can be found in the online Additional File 6 at http://www.biomedcentral.com/1471-2164/10/355/additional/.

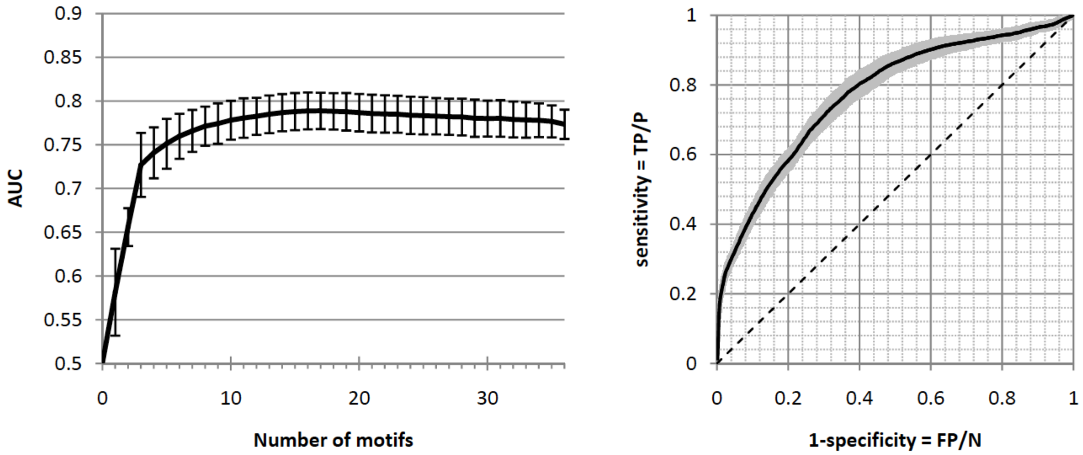


Figure 5-2. Function prediction using regulatory motifs in *T. brucei* – This figure shows inositol phosphate metabolism (KEGG:tbr00562) as an example. **(A)** The performance of our naïve Bayesian network using different numbers of motifs for prediction of inositol phosphate metabolism genes. We used a two-fold cross-validation for assessing the prediction power, where half of the dataset was used for training and the other half for validation. Cross-validation was repeated 100 times for each number of motifs, and each time the AUC (area under the curve) of the ROC curve was measured as the prediction power. Standard deviation of AUC is shown by the error bars. **(B)** The ROC curve for prediction of inositol phosphate metabolism genes using all 36 predicted motifs. Standard deviation of sensitivity is shown by the grey shaded region. The diagonal line shows the performance that would be expected if our naïve Bayesian network were not able to predict inositol phosphate metabolism genes. This classifier has a very high specificity (~99%) at sensitivities of up to 20% for this pathway.

5.3 Conclusions

The ncRNAs predicted in this study can provide candidates for experiments that are focused on understanding the functional RNA repertoire of trypanosomatids. The most interesting candidates are perhaps those that do not have characterized homologs, as they most probably represent novel ncRNA classes in *T. brucei*. Unraveling the function of these ncRNAs will help us to understand the biology of these parasites more clearly. However, it should be noted that our set of predicted ncRNAs is far from complete, as we only considered two genomes in this study. Considering a larger number of trypanosomatid genomes may reveal other ncRNAs and provide a more thorough view of the non-coding functional transcriptome of these organisms.

Prediction of gene functions based on our set of function-specific short motifs can also provide a very useful alternative to homology-based annotation methods, especially that a huge number of trypanosomatid genes are not conserved in other organisms. We anticipate that combining this method with other established systems-based function prediction approaches provides a robust method that can be applied to many genomes.

5.4 Methods

5.4.1 Identification of conserved genomic regions

Conserved genomic regions were identified using a binomial-based model [204]. Briefly, after masking LCRs using mreps, the two genomes of *T. brucei* and *L. braziliensis* were aligned using BlastZ [213]; for each window of 25 nucleotides, *N*, the number of conserved nucleotides, was determined; the probability of observing *N* conserved nucleotides out of 25 nucleotides was calculated under the null hypothesis of neutral substitution and based on binomial distribution of mutations; and regions showing evidence of negative selection were chosen. Finally, conserved regions that were less than 25 nucleotides apart were connected to each other as a single region (see [204] for detailed description of the method).

5.4.2 Identification of conserved ncRNAs

We used QRNA [205] to identify parts of the conserved genomic regions (see above) that showed patterns of conserved structural RNA elements. Long sequences were broken into smaller overlapping fragments of 80 nucleotides, each of which having 40 nucleotides overlap with its adjacent fragments. Overlapping sequences with RNA scores higher than zero were merged again, QRNA scores were recalculated, and those with final positive

RNA scores were selected as putative fRNAs. False positive rate was calculated as the fraction of conserved coding sequences that were classified as fRNA. LCRs were determined using a combination of 'mreps' and 'mdust' from TIGR, marked in the online Additional File 1 (http://www.biomedcentral.com/1471-2164/10/355/additional/) by lowercase letters. Significance of each candidate was assessed by comparing its QRNA score to a distribution obtained by randomizing *T. brucei*-*L. brazilensis* alignments: the alignment of each ncRNA candidate was considered separately, and columns with similar conservation patterns were shuffled randomly (i.e., a column containing a gap was swapped only with another gap-containing column, mismatch with mismatch, and match with match [214]). The fraction of random alignments outscoring the original alignment was considered as the *p*-value (online Additional File 1 at http://www.biomedcentral.com/1471-2164/10/355/additional/, column H).

5.4.3 Examining the candidate ncRNAs for the presence of potential pre-miRNAs.

We used a set of simple, yet powerful criteria for detection of potential miRNA precursors among our candidate fRNAs. A nucleotide sequence of length 80nt was considered a potential pre-miRNA if it could be folded into a structure with **(i)** a single stem-loop **(ii)** whose free folding energy was <= -25kcal/mol [215], **(iii)** which contained an at least 9bp-long continuous paired region with no internal loops or bulges, **(iv)** and contained no unpaired internal segment (internal loop or bulge) longer than 3 nucleotides. These criteria, while selected empirically to optimize for specificity and sensitivity, are in agreement with previous studies on pre-miRNA structure [216]. We used RNAfold from Vienna RNA package for folding the sequences.

The sensitivity and specificity of these criteria were tested on a set of 250 pre-miRNAs from 24 different organisms (Supplementary Table 5-1) and a set of 30770 randomly selected sequences from *T. brucei* genome (random sequences matching the selected criteria were considered false positives, based on which the specificity was estimated). To estimate the standard deviations of sensitivity and specificity, we performed 10 jack-knife trials in each of which one third of all sequences were randomly removed and the performance was reevaluated on the remaining two thirds. Then, all *T. brucei* genomic sequences of length 80nt that overlapped with at least one nucleotide of one of the predicted ncRNAs, as well as the reverse complements of such genomic sequences, were examined using the above criteria for presence of pre-miRNAs.

5.4.4 Finding informative regulatory elements in 5′ and 3′ UTRs

A recently developed algorithm, named FIRE, is able to identify DNA and RNA motifs that are unevenly distributed among different clusters of sequences, i.e., are overrepresented in some clusters while underrepresented in some others [168]. Here, we used FIRE to identify motifs that are unevenly distributed among different functions. Functional annotations of *T. brucei* genes were retrieved from KEGG pathway database [190]. For each pathway, we grouped the genes into two clusters based on whether they were involved in that pathway or not. Then we used FIRE to find′ 5UTR or 3′ UTR motifs that showed significant overrepresentation or underrepresentation in either of the two clusters. The sequences of mature 5and 3′ UTRs were isolated from *T. brucei* genome based on splicing site predictions reported previously [47]. The resulting motifs from different functions were collected, and duplicate motifs were removed.

The same procedure was repeated for 10 sets of randomly shuffled gene-function assignments, and the average number of motifs reported by FIRE was used as an estimate of the expected number of false positives. The expected precision was consequently calculated as $\mathrm{E}(Precision)=\mathrm{E}(TP)/P=[P\text{-}\mathrm{E}(FP)]/P$. Here, $\mathrm{E}(X)$ denotes the expected value of X, TP stands for true positive, P stands for positives (motifs detected by FIRE from actual dataset), and FP indicates false positives (motifs detected by FIRE from shuffled datasets).

It should be noted that KEGG annotations might not correspond to the precise function of the genes, since KEGG uses an automated pipeline for assigning the genes to template pathways based on homology with known proteins. However, we expect that the relationships of the genes are conserved through this procedure, e.g., if two genes are assigned to the same pathway in KEGG, they most probably have very closely related functions, even if the exact assigned functions in KEGG are not correct.

5.4.5 Function prediction using regulatory motifs

We used a naïve Bayesian network to predict gene-function assignments based on the predicted regulatory motifs in 5 and 3′ UTRs. Naïve Bayesian networks assume that the properties based on which they classify the objects are independent. Thus, the likelihood that gene *g* belongs to cluster *α* given a set of known motifs is calculated as:

$$L\left(g \in \alpha \middle| F^{M}\right) = \prod_{f_i \in F^M} L\left(g \in \alpha \middle| f_i\right)$$

Here, *M* is the set of motifs that are used for classification of *g*, and $F^M = \{f_1, f_2, \ldots, f_{|M|}\}$ where f_i is {1} if the *i*th motif is present in gene *g*, and {0} otherwise. $L(g \in \alpha \mid F^M)$ is the likelihood that *g* belongs to *α* given F^M, and $L(g \in \alpha \mid f_i)$ represents the likelihood that *g* belongs to *α* given the status of the *i*th motif in *g*, i.e., f_i. For more information about the calculation of conditional likelihoods and implementation of naïve Bayesian networks, refer to [81].

M is chosen in a way to maximize the prediction power. Briefly, motifs are selected iteratively, starting from the one that can best distinguish between *α* and *α′*. Then, at each iteration, all motifs are tested and the one whose addition to *M* results in the maximum prediction power is selected. Prediction power is assessed by the area under the ROC curve (Receiver Operating Characteristic or ROC curve plots sensitivity against false positive discovery rate). This procedure is repeated until *M* contains a predefined number of motifs. We removed paralogs prior to training and testing our naïve Bayesian network in order to avoid any biases towards duplicated UTRs.

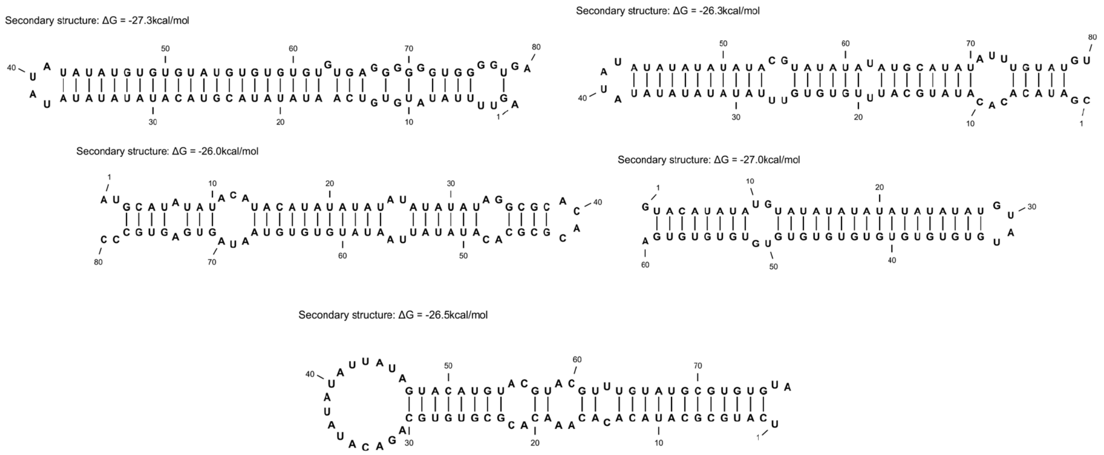
5.5 Supplementary Tables and Figures

Supplementary Table 5-1. List of microRNA sequences that were used for assessing the sensitivity of our microRNA prediction method – Accession numbers correspond to miRBase (http://microrna.sanger.ac.uk/sequences/).

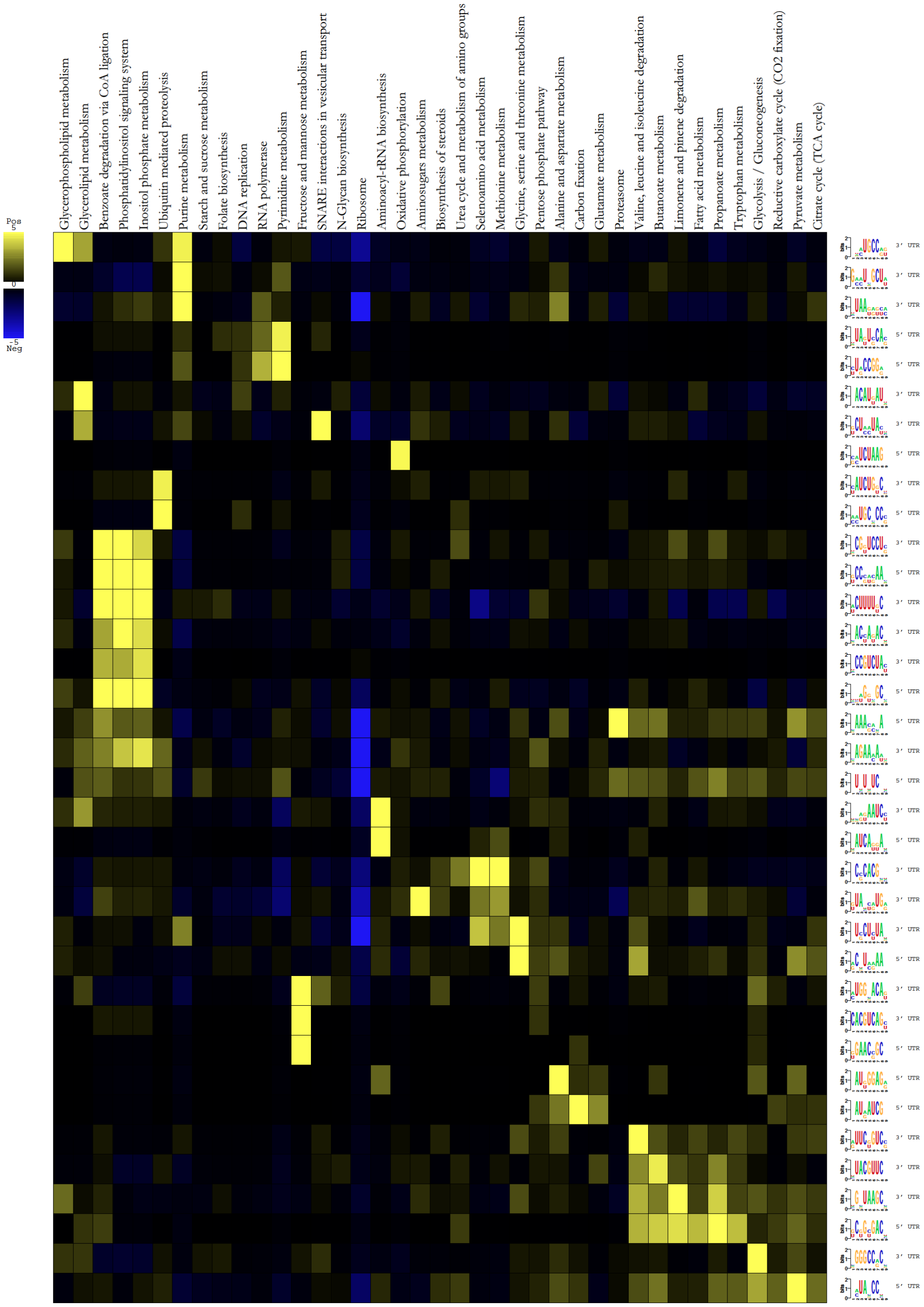
Apis mellifera		***Oikopleura dioica***		***Macaca mulatta***		***Danio rerio***	
ame-let-7	MI0005726	odi-mir-1b	MI0007134	mml-mir-16-1	MI0002958	dre-mir-1-1	MI0001878
ame-mir-2-1	MI0001589	odi-mir-124a	MI0007145	mml-mir-26a-1	MI0002646	dre-let-7a-5	MI0001862
ame-mir-9b	MI0001597	odi-mir-1469	MI0007080	mml-mir-124a-1	MI0002766	dre-let-7e	MI0001871
ame-mir-13a	MI0005730	odi-mir-1475	MI0007086	mml-mir-127	MI0002582	dre-mir-18a	MI0001900
ame-mir-79	MI0005742	odi-mir-1480	MI0007091	mml-mir-145	MI0002558	dre-mir-26a-2	MI0001925
ame-mir-100	MI0005728	odi-mir-1485	MI0007096	mml-mir-181a-2	MI0002808	dre-mir-30b	MI0001941
ame-mir-184	MI0001580	odi-mir-1489	MI0007099	mml-mir-181b-1	MI0002932	dre-mir-128-2	MI0001981
ame-mir-275	MI0005733	odi-mir-1495	MI0007108	mml-mir-181c	MI0002811	dre-mir-137-1	MI0002000
ame-mir-279	MI0005734	odi-mir-1498	MI0007120	mml-mir-221	MI0002892	dre-mir-196b	MI0002036
ame-mir-927	MI0005748	odi-mir-1502	MI0007124	mml-mir-1240	MI0006330	dre-mir-430b-10	MI0002150
Drosophila melanogaster		***Xenopus tropicalis***		***Anopheles gambiae***		***Chlamydomonas reinhardtii***	
dme-mir-1	MI0000116	xtr-let-7c	MI0004886	aga-let-7	MI0001600	cre-MIR910	MI0005703
dme-mir-2c	MI0000431	xtr-let-7e-1	MI0004907	aga-mir-12	MI0006240	cre-MIR912	MI0005705
dme-mir-31b	MI0000410	xtr-let-7e-2	MI0004909	aga-mir-34	MI0006239	cre-MIR915	MI0005708
dme-mir-284	MI0000369	xtr-let-7f	MI0004887	aga-mir-275	MI0001613	cre-MIR917	MI0005710
dme-mir-927	MI0005843	xtr-mir-7-1	MI0004790	aga-mir-281	MI0001618	cre-MIR918	MI0005697
dme-mir-960	MI0005815	xtr-mir-7-2	MI0004792	aga-mir-306	MI0006243	cre-MIR1144b	MI0006235
dme-mir-969	MI0005826	xtr-mir-7-3	MI0004791	aga-mir-989	MI0006242	cre-MIR1148	MI0006209
dme-mir-986	MI0005845	xtr-mir-27b	MI0004810	aga-mir-996	MI0006241	cre-MIR1161	MI0006222
dme-mir-1000	MI0005862	xtr-mir-98	MI0004820	aga-mir-1174	MI0006237	cre-MIR1164	MI0006225
dme-mir-1002	MI0005824	xtr-mir-196a	MI0004942	aga-mir-1175	MI0006238	cre-MIR1168	MI0006228
Caenorhabditis elegans		***Gallus gallus***		***Mus musculus***		***Physcomitrella patens***	
cel-let-7	MI0000001	gga-let-7b	MI0001172	mmu-mir-1-2	MI0000652	ppt-MIR156c	MI0005654
cel-mir-1	MI0000003	gga-mir-215	MI0001203	mmu-mir-29b-2	MI0000712	ppt-MIR160d	MI0005658
cel-mir-36	MI0000007	gga-mir-16-2	MI0001222	mmu-mir-30d	MI0000549	ppt-MIR160g	MI0005903
cel-mir-41	MI0000012	gga-mir-27b	MI0001274	mmu-mir-98	MI0000586	ppt-MIR166g	MI0005910
cel-mir-51	MI0000022	gga-mir-187	MI0001193	mmu-mir-101b	MI0000649	ppt-MIR390a	MI0003494
cel-mir-65	MI0000036	gga-mir-7-3	MI0001269	mmu-mir-129-1	MI0000222	ppt-MIR538b	MI0003511
cel-mir-70	MI0000041	gga-mir-181b-1	MI0001219	mmu-mir-410	MI0001161	ppt-MIR1029	MI0005974
cel-mir-243	MI0000319	gga-mir-301	MI0001240	mmu-mir-467h	MI0006302	ppt-MIR1056	MI0006015

cel-mir-784	MI0005184	gga-mir-20b	MI0001517	mmu-mir-669b	MI0004666	ppt-MIR1216	MI0004715
cel-mir-1818	MI0007980	gga-mir-146a	MI0001235	mmu-mir-669e	MI0006300	ppt-MIR1223c	MI0005952
Homo sapiens		***Canis familiaris***		***Rattus norvegicus***		***Arabidopsis thaliana***	
hsa-mir-7-3	MI0000265	cfa-mir-1-1	MI0008060	rno-let-7f-2	MI0000834	ath-MIR156e	MI0000182
hsa-mir-135a-2	MI0000453	cfa-mir-26a-2	MI0007990	rno-mir-19b-2	MI0000848	ath-MIR161	MI0000193
hsa-mir-181d	MI0003139	cfa-mir-212	MI0008155	rno-mir-103-1	MI0000888	ath-MIR167b	MI0000209
hsa-mir-320b-2	MI0003839	cfa-mir-365-1	MI0001657	rno-mir-138-2	MI0000911	ath-MIR169e	MI0000979
hsa-mir-496	MI0003136	cfa-mir-365-1	MI0001657	rno-mir-181b-2	MI0000927	ath-MIR172e	MI0001089
hsa-mir-548k	MI0006354	cfa-mir-365-2	MI0001647	rno-mir-222	MI0000962	ath-MIR395f	MI0001012
hsa-mir-567	MI0003573	cfa-mir-429	MI0001644	rno-mir-380	MI0006141	ath-MIR781	MI0005111
hsa-mir-633	MI0003648	cfa-mir-448	MI0001640	rno-mir-503	MI0003555	ath-MIR827	MI0005383
hsa-mir-1197	MI0006656	cfa-mir-449	MI0001651	rno-mir-742	MI0006161	ath-MIR840	MI0005396
hsa-mir-1302-1	MI0006362	cfa-mir-450a	MI0001655	rno-mir-878	MI0006120	ath-MIR860	MI0005437
Schmidtea mediterranea		***Monodelphis domestica***		***Bos taurus***		***Brassica napus***	
sme-let-7a	MI0005122	mdo-let-7a-1	MI0005360	bta-let-7c	MI0005454	bna-MIR166b	MI0006475
sme-lin-4	MI0005125	mdo-let-7b	MI0005351	bta-mir-29a	MI0004733	bna-MIR167a	MI0006471
sme-mir-2c	MI0005134	mdo-mir-10b	MI0005274	bta-mir-30d	MI0004747	bna-MIR168	MI0006470
sme-mir-7b	MI0005137	mdo-mir-19b	MI0005358	bta-mir-127	MI0005008	bna-MIR169a	MI0006457
sme-mir-12	MI0005141	mdo-mir-27b	MI0005367	bta-mir-181c	MI0005032	bna-MIR169g	MI0006463
sme-mir-79	MI0005153	mdo-mir-107	MI0005286	bta-mir-200a	MI0005037	bna-MIR169m	MI0006469
sme-mir-184	MI0005163	mdo-mir-181b	MI0005344	bta-mir-215	MI0005016	bna-MIR171d	MI0006453
sme-mir-747	MI0005166	mdo-mir-214	MI0005319	bta-mir-345	MI0005019	bna-MIR390a	MI0006447
sme-mir-752	MI0005178	mdo-mir-340	MI0007268	bta-mir-365	MI0005465	bna-MIR390c	MI0006449
sme-mir-756	MI0005182	mdo-mir-365	MI0005326	bta-mir-497	MI0005467	bna-MIR397b	MI0006446
Populus trichocarpa		***Oryza sativa***		***Triticum aestivum***		**Kaposi sarcoma-associated herpesvirus**	
ptc-MIR156i	MI0002192	osa-MIR439a	MI0001691	tae-MIR171	MI0006175	kshv-miR-K12-2	MI0002476
ptc-MIR162a	MI0002209	osa-MIR444c	MI0006975	tae-MIR408	MI0006177	kshv-mir-K12-12	MI0004987
ptc-MIR162b	MI0002210	osa-MIR444f	MI0006978	tae-MIR1117	MI0006179		
ptc-MIR164b	MI0002213	osa-MIR535	MI0003505	tae-MIR1118	MI0006180	**Mouse gammaherpesvirus 68**	
ptc-MIR168a	MI0002243	osa-MIR809h	MI0005228	tae-MIR1120	MI0006182	mghv-mir-M1-4	MI0001672
ptc-MIR171f	MI0002282	osa-MIR810a	MI0005229	tae-MIR1124	MI0006186		
ptc-MIR171n	MI0007034	osa-MIR819d	MI0005255	tae-MIR1125	MI0006187	**Rhesus lymphocryptovirus**	
ptc-MIR478d	MI0002373	osa-MIR819h	MI0005259	tae-MIR1125	MI0006187	rlcv-mir-rL1-3	MI0003739
ptc-MIR1449	MI0007049	osa-MIR1430	MI0006970	tae-MIR1134	MI0006196	rlcv-mir-rL1-9	MI0003745
ptc-MIR1446a	MI0007042	osa-MIR1436	MI0007022	tae-MIR1137	MI0006199		
Ciona intestinalis		**Epstein barr virus**		**Rhesus monkey rhadinovirus**			
cin-let-7a-1	MI0007149	ebv-mir-BART5	MI0003727	rrv-miR-rR1-3	MI0005720		

cin-let-7d	MI0007153	ebv-mir-BART16	MI0004989
cin-mir-31	MI0007156		
cin-mir-34	MI0007158	**Herpes simplex virus 1**	
cin-mir-92a	MI0007160	hsv1-mir-H1	MI0004730
cin-mir-126	MI0007166		
cin-mir-141	MI0007168	**Human cytomegalovirus**	
cin-mir-153	MI0007169	hcmv-mir-US4	MI0003687
cin-mir-200	MI0007175		
cin-mir-219	MI0007178		


Secondary structure: ΔG = -27.3kcal/mol
Secondary structure: ΔG = -26.3kcal/mol
Secondary structure: ΔG = -26.0kcal/mol
Secondary structure: ΔG = -27.0kcal/mol
Secondary structure: ΔG = -26.5kcal/mol

Supplementary Figure 5-1. miRNA-like predicted ncRNAs – Candidate ncRNAs whose predicted secondary structures match our criteria for miRNA prediction are shown in this figure. It can be seen that their sequences mostly consist of AU repeats, rendering them unlikely candidates for being miRNA.



Supplementary Figure 5-2. Function-specific regulatory motifs that were identified in 5′ and 3′ UTRs of *T. brucei* – Each row represents one motif, while each column stands for one function. Overrepresentation of a motif in a function is indicated by a yellow square, while underrepresentation is shown by blue. The probabilities of overrepresentation or underrepresentation were calculated based on hypergeometric distribution assumption and are shown here by the color gradient on log scale.

6 Predicting molecular functions of non-conserved proteins using a comprehensive catalogue of function-specific short sequence signatures

In chapters 3-5 we discussed novel methods for homology-independent sequence-based prediction of biological processes and pathways. However, the pathway to which a protein belongs is only partially informative about the exact function of a protein. Knowing the molecular function(s) of a protein in addition to its associated biological process(es), one can more precisely predict its biological role. In this chapter, we describe a catalogue of more than 4800 function-specific short protein motifs that can be used for homology-independent prediction of protein molecular functions. We show that this catalogue, which is obtained from analysis of all molecular functions in the GO database, represents known as well as novel protein functional sites such as enzyme active sites and ligand-binding pockets. We then present classifiers that use these motifs to predict molecular functions of proteins, and describe their application in annotating the uncharacterized proteins of *Trypanosoma brucei*. The function-specific short protein motif catalogue that is presented here, along with the provided software, serves as a new resource for functional annotation of uncharacterized proteins in different organisms.

6.1 Background

The massive amount of available annotated nucleic acid and amino acid sequences has provided an unprecedented source of information for functional annotation of new genes and proteins. The majority of current automated methods for function prediction use homology-based approaches, in which a query sequence is searched against a database of sequences or patterns with known functions, and a predicted function is then assigned to the query sequence based on its similarity to the database entries [217]. However, along with publication of new sequences emerge novel families of proteins that share little similarity with functionally characterized known proteins [218, 219]. Furthermore, proteins of evolutionary distant organisms are too diverged to be reliably aligned to characterized proteins of well-studied organisms. These factors severely limit the applicability of homology-based methods, leaving large portions of newly released sequences uncharacterized.

While homology-dependent methods are based on identification of conserved features among evolutionary related proteins, other methods rely on convergent evolution of features in proteins of similar functions. Examples of such features include sequence length, physiochemical properties, and amino acid composition. It has been shown that such features can be integrated in order to increase the performance of protein classifiers [220]. However, their specificity is a matter of concern, given that they are mostly shaped by more general factors such as subcellular localization, rather than by the specific molecular function of the protein.

Function-specific short sequence signatures constitute a group of features that may have been formed by either convergent or divergent evolution. Sequence signatures that are formed by convergent evolution mostly function as mediators of protein-protein interactions, signal peptides for protein sorting, and recognition sites for protein-modifying enzymes [221]. The most notable of short sequence signatures that are conserved during divergent evolution of homologous proteins are enzyme active sites, ligand-binding residues, and signatures that represent structurally/functionally critical regions of protein domains [222]. Unlike long sequence profiles, short sequence signatures can be detected even in proteins that are only remotely related to characterized proteins with known molecular functions. However, because of their small size, they can sporadically occur in unrelated sequences, overshadowing the true instances by a large number of false positives.

Whether created through convergent evolution or conserved during divergent evolution, function-specific short sequence signatures can lead to new approaches toward prediction of functions of non-conserved or less-conserved proteins. A few specialized databases are allocated to short protein motifs, including The Eukaryotic Linear Motif Resource [223] and PROSITE [224]. Also, several methods have been developed for computational identification of short protein motifs, particularly from protein interaction networks [225-227] and proteomics data [228]. However, these studies have been mostly concerned with identification of short motifs rather than their application in predicting molecular functions of proteins. In addition, lack of a comprehensive analysis of available functional data for identification of function-specific short protein motifs is apparent.

Here, we present a thorough analysis of available protein sequences in the Gene Ontology database in order to identify function-specific short sequence signatures. We explore the application of these signatures in predicting molecular functions of proteins, and show that despite their short profiles, they can effectively predict protein functions if they are appropriately integrated. We then apply our method to annotate the proteins of *Trypanosoma brucei*, a deadly parasite of the trypanosomatid clade, and demonstrate that this approach can predict the functions of proteins for which homology-based methods fail. This analysis provides a novel resource for functional annotation of newly sequenced genomes.

6.2 Materials and Methods

Gold standard datasets – We extracted 27802 protein-molecular function associations from GO database, considering only annotations that were supported by the evidence code IDA (Inferred from Direct Assay). Sequences associated with these proteins were also extracted from GO. We ensured that our dataset did not contain any proteins from *Trypanosoma brucei* by removing proteins that could be aligned to at least one *T. brucei* protein with≥95% identity. This step was necessary, as we would later validate our classifiers using *T. brucei* annotations. For each target GO term, positive and negative gold standard sets were determined considering the directed acyclic graph nature of GO database, as previously described [229]. Gold standard sets were further filtered to remove homologous proteins. Filtering homologous proteins was necessary to prevent the subsequent analysis from identifying abundant signals of homologous regions, which may overshadow the weaker signatures from less conserved regions or signatures of convergent evolution. We used BLAST-P to find homologous pairs with E-value ≤1. If two homologs happened to be both in the positive gold standard set or both in the negative gold standard set for a particular GO term, one of them was randomly chosen to be removed, ensuring that no protein pairs in either of the positive or negative sets had homologous regions. Overall, 13986 protein sequences and 3891 GO terms were analyzed.

Identification of function-specific short protein motifs – We developed an algorithm, called HyperMotif, which searches for degenerate short protein patterns that are significantly over-represented in at least one GO term. Briefly, HyperMotif creates an exhaustive list of degenerate short motifs by running a sliding window of size ≤6 through the sequences of all input proteins, and considers degenerate versions of each of the obtained peptides by replacing different amino acids with the symbol X (which represents a completely degenerate site). Each degenerate motif is then examined for its overlap with each GO category if that motif has at least four non-degenerate residues, and if at least one of the proteins within that GO category has the motif. The significance of overlap is evaluated using hypergeometric test (see Supplementary Figure 6-1). False discovery rate (FDR) is then calculated based on the obtained *p*-values and the number of motif-category comparisons. In this study, we chose to accept motif-category associations with FDR ≤0.1.

HyperMotif, accompanied by several accessory programs that allow subsequent motif analysis and protein function prediction, is available at http://webpages.mcgill.ca/staff/Group2/rsalav/web/Software/HyperMotif/index.htm.

Identification of PROSITE patterns that match the discovered motifs – To evaluate the ability of HyperMotif in finding functional protein motifs, we compared the patterns reported by HyperMotif to the patterns in PROSITE database. Matching patterns were identified based on the extent of overlap of their instances. Briefly, the 13986 protein sequences from GO database were searched for instances of PROSITE patterns, excluding profiles as well as frequently matching (unspecific) patterns. For a HyperMotif pattern of length l, we determined the total number of sliding windows of length l in the sequence database, the number of sliding windows that overlapped at least one residue of at least one instance of a specific PROSITE pattern, the number of sliding windows that actually matched the HyperMotif pattern, and the number of sliding windows that met both of the last two criteria. Then, these numbers were used to calculate the p-value of overlap of the HyperMotif pattern and the PROSITE pattern based on hypergeometric test. We chose a stringent Bonferroni-corrected p-value threshold of 0.005.

Prediction of molecular functions – We developed an algorithm to integrate the motifs from HyperMotif using naïve Bayesian classifiers in order to predict protein molecular functions. First, motif-category associations that are still significant even after removal of any single protein from the gold standard dataset are identified. This step removes motifs that would not show up if a leave-one-out cross-validation method was used for motif identification. Then, for each category, a specific naïve Bayesian classifier is created using the motifs that significantly overlap that category in all leave-one-out tests, and a likelihood of association with that category is assigned to each protein based on the presence or absence of the motifs [see Ref. [81] for a description of naïve Bayesian networks]. The performance of the naïve Bayesian classifier is evaluated by leave-one-out cross validation, and a likelihood threshold is chosen in order to obtain a precision of 0.8 – precision is defined as TP/(TP+FP), where TP is the number of true positives and FP is the number of false positives. A p-value is also assigned to each classifier, indicating the chance that a random classifier would result in the same or better sensitivity at precision of 0.8 – sensitivity is defined as TP/(TP+FN), where FN is the number of false negatives. The p-value is calculated using hypergeometric test by evaluating the overlap of the set of positive proteins from gold standard dataset and the set of proteins whose likelihood scores are greater than the likelihood threshold (Supplementary Figure 6-2). The obtained classifiers were then applied to the genome of *T. brucei* to predict the molecular functions of *T. brucei* proteins. Predictions were compared to GO molecular functions that are known for *T. brucei* proteins according to TriTrypDB version 2.4 [230], accepting all evidence codes except for IC, IRD, IKR, IBD, IBA, NAS, and ND. For catalytic activity, predictions were improved by considering the number of transmembrane helices that are predicted for each protein using SOSUI [231].

6.3 Results and Discussion

***HyperMotif identifies short protein signatures of different molecular functions* –** Analysis of the complete GO database using HyperMotif identified 6496 associations between 4883 short protein motifs and 414 GO molecular functions at $\text{FDR} \leq 0.1$. Of these associations, 513 are significant at $\text{FDR} \leq 0.01$. Many of the discovered protein motifs match known functional sites. For example, GXGKT, which obtained the most significant *p*-value for association with nucleoside-triphosphatase activity, matches the sequence of Walker A motif, a well-known ATP-binding motif that is found in many ATP hydrolyzing enzymes [232]. The next highest-scoring associations belong to several proline-rich motifs that are associated with transcription regulator activity, congruent with the long-known *trans*-activation of transcription by proline-rich domains [233]. The HEXGH motif also closely resembles the HEXXH motif which forms part of the metal-binding site of metalloproteases [234], congruent with the association found by HyperMotif between HEXGH motif and metallopeptidase activity.

To systematically compare HyperMotif patterns with our current knowledge of short protein signatures, we searched for PROSITE patterns that matched the motifs found by HyperMotif. The PROSITE database [224] contains many short patterns (Supplementary Figure 6-3) that represent protein functional sites. We identified a total of 469 matches between 420 HyperMotif patterns and 181 PROSITE patterns (online Supplementary Table 1 at http://webpages.mcgill.ca/staff/Group2/rsalav/web/Suppl/20110810/Tables/Supplementary%20Table%20S1.xlsx). This extent of overlap validates a large number of motifs identified by HyperMotif as known functional protein sites. It is not surprising to observe that in several cases one PROSITE pattern matches more than one HyperMotif pattern, given that HyperMotif examines an exhaustive set of degenerate motifs that may be closely similar to each other. For example, two motifs PGXSG and PGXXGP match the same region of the PROSITE entry for EGF-like domain signature 1. To examine the level of redundancy of patterns that were discovered by HyperMotif, we measured the overlap among HyperMotif patterns using the same methodology that we employed for measuring the overlap of HyperMotif patterns and PROSITE patterns. We found that a relatively small fraction of HyperMotif patterns overlap each other (Figure 6-1A), meaning that most of these short patterns are distinct functional sites.

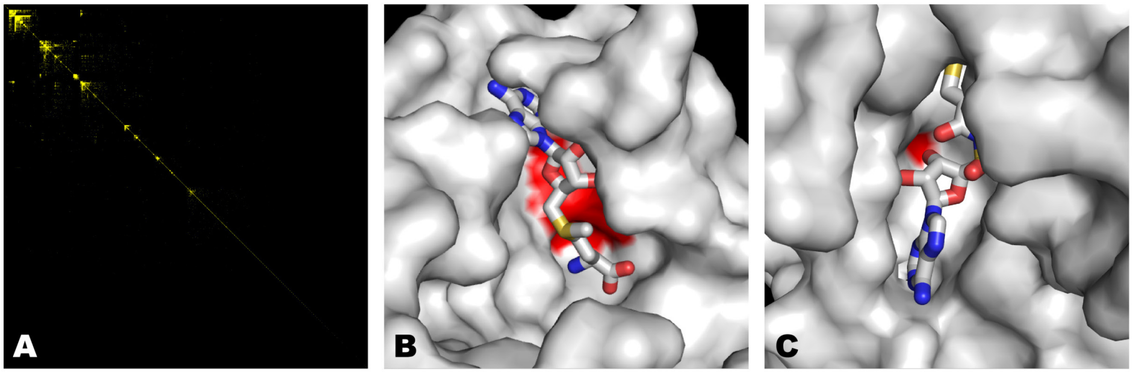


Figure 6-1. HyperMotif identifies many non-redundant motifs representing functional protein sites – (A) There is not much overlap among instances of different HyperMotif patterns. Each row and each column represent one HyperMotif pattern. A yellow dot represents two motifs whose instances significantly overlap each other (redundant motifs). **(B)** A close-up view of Dengue-3 NS5 methyltransferase bound to the substrate S-adenosyl methionine (PDB accession number 3P97). The methyltransferase-specific motif DXGCG is highlighted in red, and the substrate S-adenosyl methionine is shown as sticks. **(C)** Mycobacterium smegmatis MshC in complex with 5'-O-(N-(L-cysteinyl)-sulfamoyl)-adenosine (PDB accession number 3C8Z). The red patch highlights the accessible surface of the motif DIXXGG which is specific to adenyl-ribonucleotide binding proteins.

More than 4400 of the motifs found by HyperMotif do not have a match in PROSITE, many of which may represent novel function-specific motifs. For example, HyperMotif found a highly significant overlap between the set of proteins that contain the motif DXGCG and the set of proteins that have methyltransferase activity – about half of the proteins that contain this motif are methyltransferases ($p<1\times10^{-6}$). Analysis of the structure of known methyltransferases that contain this motif showed that this motif indeed directly interacts with the methyl donor substrate of these enzymes, i.e. S-adenosyl-methionine (Figure 6-1B). Another example is a novel DIXXGG motif, which was found to be highly enriched among adenyl-ribonucleotide binding proteins ($p<1\times10^{-5}$). We found that this motif interacts with the ribose or ribitol of adenyl-ribonucleotide derivatives (Figure 6-1C).

Intriguingly, neither the popular motif-discovery tool MEME [235] nor the more recent FIRE-pro [228] were able to identify the adenyl-ribonucleotide binding motif DIXXGG. Instead of the DIXXGG motif, FIRE-pro reported several highly degenerate motifs that we were not able to correlate with the available protein structures, except for a Walker A-like motif which was also discovered by MEME (Supplementary Figure 6-4). FIRE-pro also failed to identify the methyltransferase-specific motif DXGCG. In comparison, MEME was able to discover the DXGCG motif with an E-value of 7.8×10^{-9} (Supplementary Figure 6-5) based on the sequences of proteins with methyltransferase activity. However, MEME also identified several motifs with similar E-values in equally sized random samples of proteins that did not have any common functions – in eight out of 20 random protein sets MEME identified at least one motif with E-value $\leq7.8\times10^{-9}$. This is while at a p-value of 5.5×10^{-9} HyperMotif did not identify any motifs in protein sets whose molecular functions were randomly shuffled (5.5×10^{-9} corresponds to the p-value of association of DXGCG and methyltransferase activity). These few examples suggest a higher specificity and sensitivity for HyperMotif in comparison to MEME and FIRE-pro. Furthermore, the ability of HyperMotif to correctly model the directed acyclic graph of Gene Ontology makes it more suitable for identification of motifs from ontology-based categories. In the next section, we address the question as to whether we have captured enough information in the form of these motifs to predict the molecular functions of proteins based on their presence and absence.

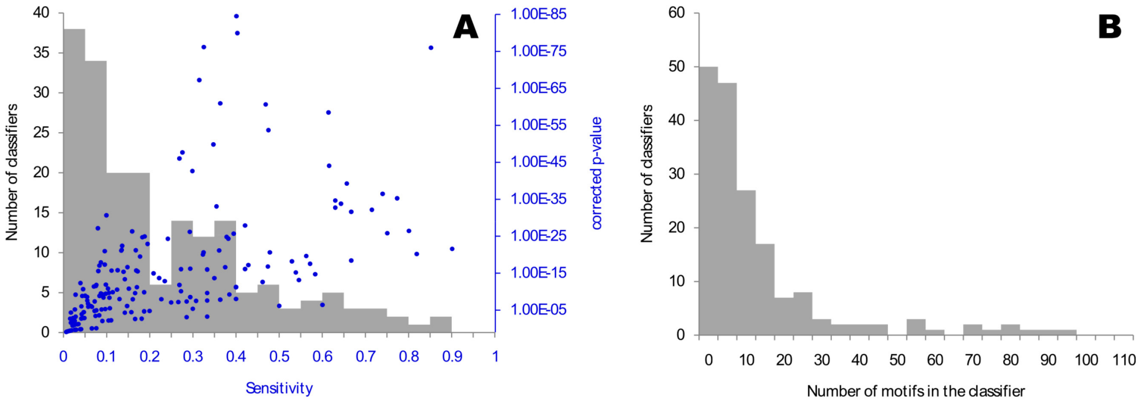


Figure 6-2. HyperMotif patterns can predict molecular functions of different proteins – (A) At 0.8 precision, different GO molecular functions are predicted with different sensitivities. While the sensitivity of predicting many molecular functions is less than 0.5, some other molecular functions can be predicted with sensitivities as high as 0.9. Nonetheless, in most cases the obtained sensitivities are far better than would be expected by random. **(B)** Different classifiers use different numbers of motifs to predict the proteins that are associated with their corresponding molecular functions. These motifs are integrated into a naïve Bayesian classifier, as described in the text.

Short motifs are able to predict protein molecular functions – By combining the HyperMotif patterns using naïve Bayesian classifiers, we were able to predict 192 GO molecular functions at 0.8 precision (Figure 6-2A and online Supplementary Table 2 at http://webpages.mcgill.ca/staff/Group2/rsalav/web/Suppl/20110810/Tables/Supplementary%20Table%20S2.xlsx). Of these, the sensitivity of 173 classifiers was significantly higher than what would be expected from a random classifier at 0.8 precision (Bonferroni corrected p-value ≤0.05). In many cases, only a few signature motifs were enough to obtain 0.8 precision: 50 out of 173 classifiers used five or fewer motifs to identify proteins that belonged to their respective molecular functions, and more than 70% of the classifiers used at most 15 short motifs (Figure 6-2B and online Supplementary Table 3 at http://webpages.mcgill.ca/staff/Group2/rsalav/web/Suppl/20110810/Tables/Supplementary%20Table%20S3.xlsx). In contrast to HyperMotif patterns, PROSITE patterns were able to predict no more than four molecular functions at 0.8 precision; glycosyltransferase activity and pentosyltransferase activity could be both predicted based on the PROSITE entry for purine/pyrimidine phosphoribosyl transferases signature, cysteine-type peptidase activity could be predicted using either of the PROSITE entries ubiquitin carboxyl-terminal hydrolases family 2 signatures 1 or 2, and cysteine-type endopeptidase inhibitor activity could be predicted based on the presence of cysteine proteases inhibitors signature. This clearly shows the utility of a comprehensive catalogue of function-specific short motifs similar to what we are reporting here. Furthermore, this analysis suggests that a much wider range of functions are covered by HyperMotif patterns in comparison to existing databases.

Exclusion of GO annotations that are not inferred from direct assay has enabled us to define a highly reliable and conservative set of positive gold standards for each GO molecular function. Obtaining a reliable negative gold standard set, however, is not as trivial, as we may inadvertently include some potential positives in the negative gold standard set. A possible source of error in constructing an accurate negative gold standard set is the incomplete knowledge that we have about the different functions that proteins may have. For example, our analysis predicts a novel interaction between phosphodiesterase 3B (Pde3B) and beta-catenin. While this prediction is considered a false positive when estimating the error rate of our classifier for the beta-catenin binding category, it may actually describe a previously unidentified function of Pde3B. This is particularly likely as both Pde3B and beta-catenin are known to be activated by insulin [236, 237], indicating a possible functional interaction (cross-talk) between the two proteins (although the current signaling pathways that are proposed for insulin-mediated activation of these two proteins are different). This incomplete knowledge has probably resulted in the underestimation of the precision of our classifiers, and perhaps the performance of HyperMotif patterns in prediction of molecular functions is better than what we have estimated here. This problem is more obvious when we try to evaluate the predictions of our classifiers in poorly annotated genomes, as we will discuss in the next section.

Prediction of protein molecular functions in Trypanosoma brucei – As described in the previous section, we have validated the naïve Bayesian classifiers by performing leave-one-out cross-validation – i.e. one protein is removed from the training set, the parameters of the naïve Bayesian network are trained on the remaining proteins, and the function of the left-out protein is then predicted using the naïve Bayesian network. However, these classifiers were built using motifs that were identified from the whole gold standard set, raising the possibility that the cross-validation that we performed was 'contaminated' [238]. As described in the methods section, we tried to address this issue by filtering out motifs that would not be significant if any of the gold standard proteins was removed. In other words, features that were used by the naïve Bayesian classifiers were 'robust', and would be used even if leave-one-out cross-validation were carried out from the motif-finding step. To further validate our classifiers and examine their utility in predicting molecular functions, we applied them to the genome of *T. brucei*, an early diverged eukaryote whose proteins were excluded from the training set during motif finding and Bayesian parameter selection.

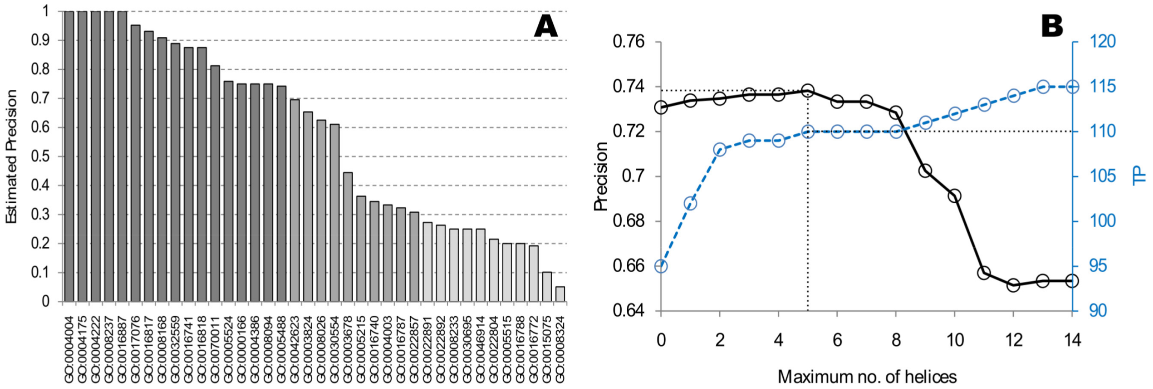


Figure 6-3. HyperMotif patterns can be used to predict several molecular functions in *T. brucei* proteins –(A) Molecular functions are grouped based on the estimated precision of their corresponding predictions in *T. brucei*, represented here by different shadings. This figure only shows the molecular functions for which at least one known true positive was among the predictions. **(B)** By considering the number of potential helices, proteins with catalytic activity can be predicted at higher precision. The highest precision is obtained when proteins that have more than five helices are filtered out. At this precision, the number of true positives is barely reduced, meaning a negligible loss of sensitivity.

T. brucei is a human parasite of the trypanosomatid clade which is responsible for the death of several thousand individuals per year. The genome sequence of this organism harbors a large number of uncharacterized proteins with no predicted functions, primarily because of the lack of similarity between their sequences and the sequences of characterized proteins or protein domains from other organisms. This evolutionary distance has rendered *T. brucei* an appropriate model to test homology-independent methods of protein function prediction.

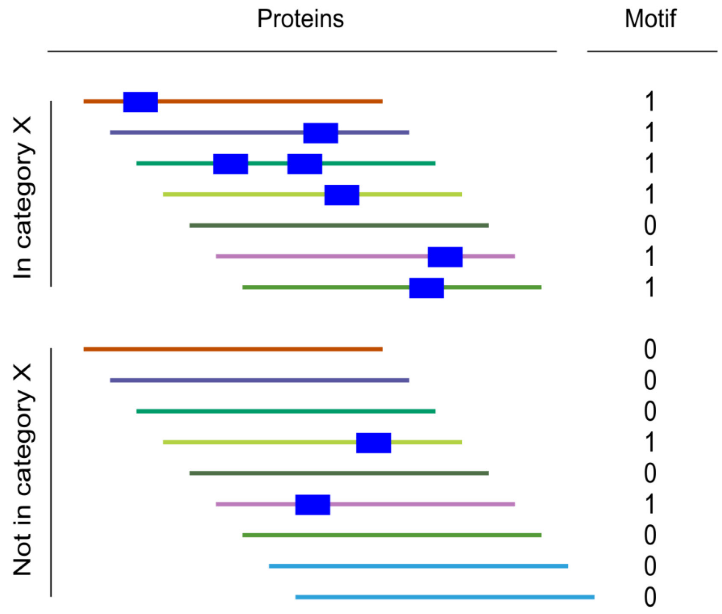
Using the naïve Bayesian classifiers that were originally trained on GO database excluding *T. brucei* proteins, we were able to make a total of 4666 predictions for *T. brucei* proteins, consisting of 2675 predictions for previously uncharacterized proteins. We evaluated the performance of each classifier (for each molecular function) separately using available molecular function annotations of *T. brucei*. As shown in Figure 6-3A and online Supplementary Table 4 (http://webpages.mcgill.ca/staff/Group2/rsalav/web/Suppl/20110810/Tables/Supplementary%20Table%20S2.xlsx), HyperMotif patterns were able to identify 17 different categories of molecular functions with precisions≥0.7 (high quality predictions) and an additional 10 categories with precisions between 0.3 and 0.7 (moderate quality predictions). It should be noted that these estimates suffer from the same errors that we discussed above for performance evaluation, i.e. we may underestimate the number of true positives and overestimate the number of false positives. This is especially the case for a poorly characterized genome as that of *T. brucei*. This uncertainty became more pronounced when we realized that the precision of predictions for 53 categories could not be estimated, because none of the *T. brucei* proteins that we assigned to those categories had been characterized previously. Therefore, although we have grouped our predictions into high quality, moderate quality and low quality categories, this categorization may simply be an artifact of our inability in detecting true positive predictions. Yet, high quality and moderate quality sets contain 329 and 616 predictions for previously uncharacterized proteins, respectively, clearly indicating that function-specific short motifs can predict the functions of a large number of proteins for which homology-based approaches fail.

Since GO categories have different levels of information, the predictions that are made are not informative at the same level; while some predictions are highly informative of the function of the protein, such as those for ATP-dependent RNA helicase activity (GO:0004004), some others are very general and cover a wide range of activities, such as the predictions for catalytic activity (GO:0003824). Nonetheless, the low-information predictions can also contribute to the annotation of *T. brucei* proteins and guide future studies. For example, we have predicted a total of 152 previously uncharacterized proteins to have catalytic activity with a precision of 0.65. This represents a repertoire of *T. brucei* enzymes that can potentially fill the many "pathway holes" [239] that are present in the metabolism map of this organism. The precision of predictions for these potential enzymes can be greatly improved by incorporating orthogonal information, such as protein structure. We found that there is a much higher chance of having catalytic activity if the protein has at most five transmembrane helices. Among proteins that were predicted based on short motifs to have catalytic activity, only five true positives had more than five transmembrane helices, while 110 true positives had five or fewer transmembrane helices. On the other hand, 22 false positives contained more than five helices. Therefore, simply by filtering out the proteins that had more than five helices, we were able to remove a considerable number of false positives and improve the precision of our predictions to 0.74 without a notable decrease in sensitivity (Figure 6-3B). This improved set of predicted enzymes contains 145 previously uncharacterized proteins (online Supplementary Table 5 at http://webpages.mcgill.ca/staff/Group2/rsalav/web/Suppl/20110810/Tables/Supplementary%20Table%20S5.xlsx).

The function-specific short motif catalogue that we presented in this study can be expanded by considering motifs of various lengths and a wider range of possibilities for the number of degenerate sites. Furthermore, the motif definitions can be refined by using more detailed regular expressions such as those used by PROSITE; in contrast to PROSITE, HyperMotif only considers non-degenerate or completely degenerate residues at the moment. In addition, classifiers that use more sophisticated structures than naïve Bayesian networks may result in better classification performance, as the redundancy of the motif catalogue can be more accurately modeled in the structure of those classifiers.

Nonetheless, our short motif catalogue should be readily able to predict the molecular functions of many previously uncharacterized proteins in different organisms. These motifs will also guide biochemical studies by pointing to the functional sites that are directly involved in the predicted function of each protein.

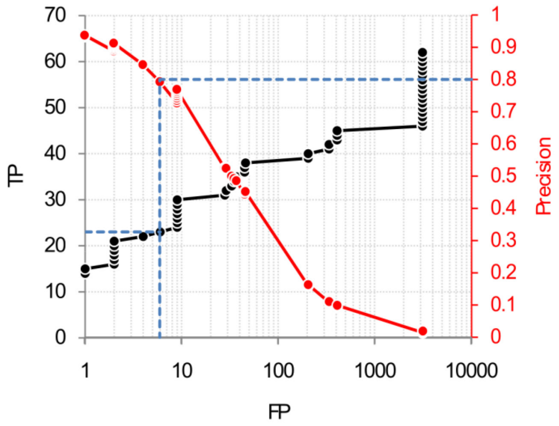
6.4 Supplementary Figures



	Have the motif	Do not have the motif	Total
In category X	6	1	7
Not in category X	2	7	9
Total	8	8	16

$$P(Y \geq 6) = \sum_{y=6}^{7} \frac{\binom{7}{y}\binom{9}{8-y}}{\binom{16}{8}} = 0.02$$

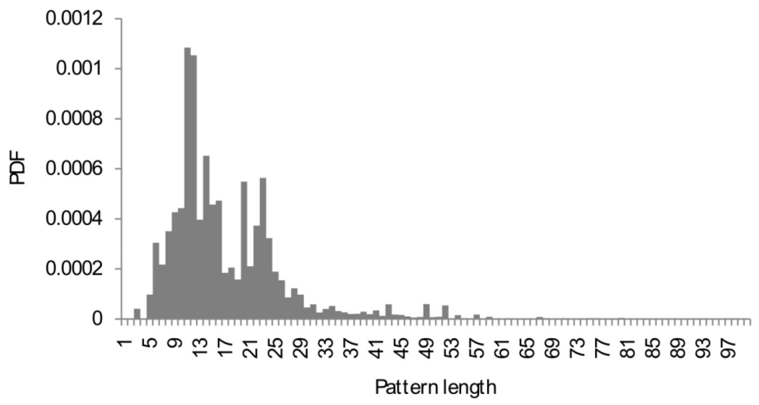
Supplementary Figure 6-1. Calculating the p-value of association of a particular motif with a particular protein category – The overlap of the set of proteins that are in category X with the set of proteins that have the motif M is evaluated by a hypergeometric test. This toy example shows a gold standard positive set of seven proteins and a gold standard negative set of 9 proteins (i.e. seven proteins have the molecular function of interest and nine proteins do not have that molecular function). Of these 16 proteins, eight have at least one instance of motif M, six of which being in the positive gold standard set (the motif instances are shown by the blue squares). On the right side of this figure, the corresponding contingency table is shown, along with the calculation of the *p*-value based on assumption of hypergeometric distribution.



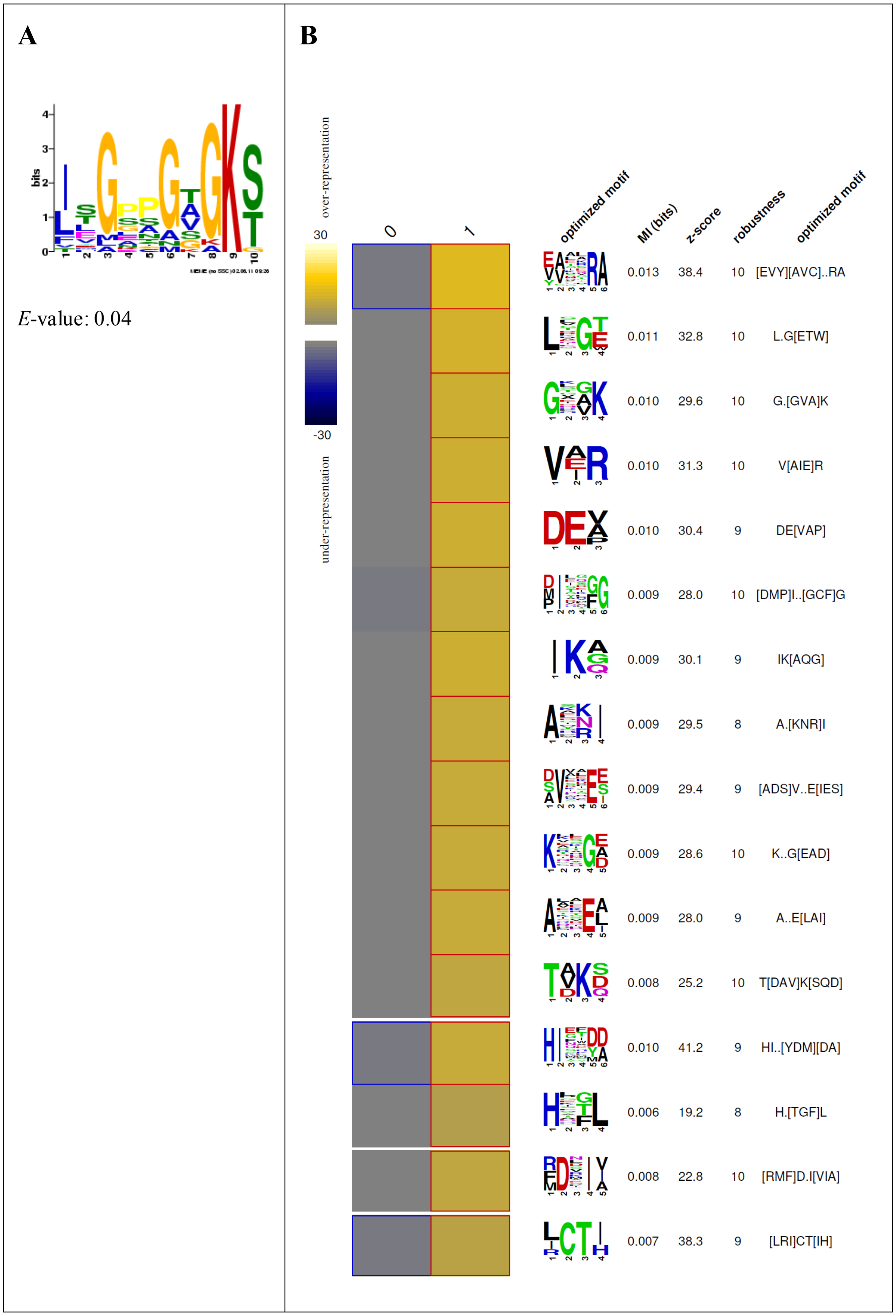
	TP	FP	Total
Above 0.8 precision cutoff	23	6	29
Below 0.8 precision cutoff	39	3132	3171
Total	62	3138	3200

$$P(Y \geq 23) = \sum_{y=23}^{29} \frac{\binom{29}{y}\binom{3171}{62-y}}{\binom{3200}{62}} = 4.55 \times 10^{-36}$$

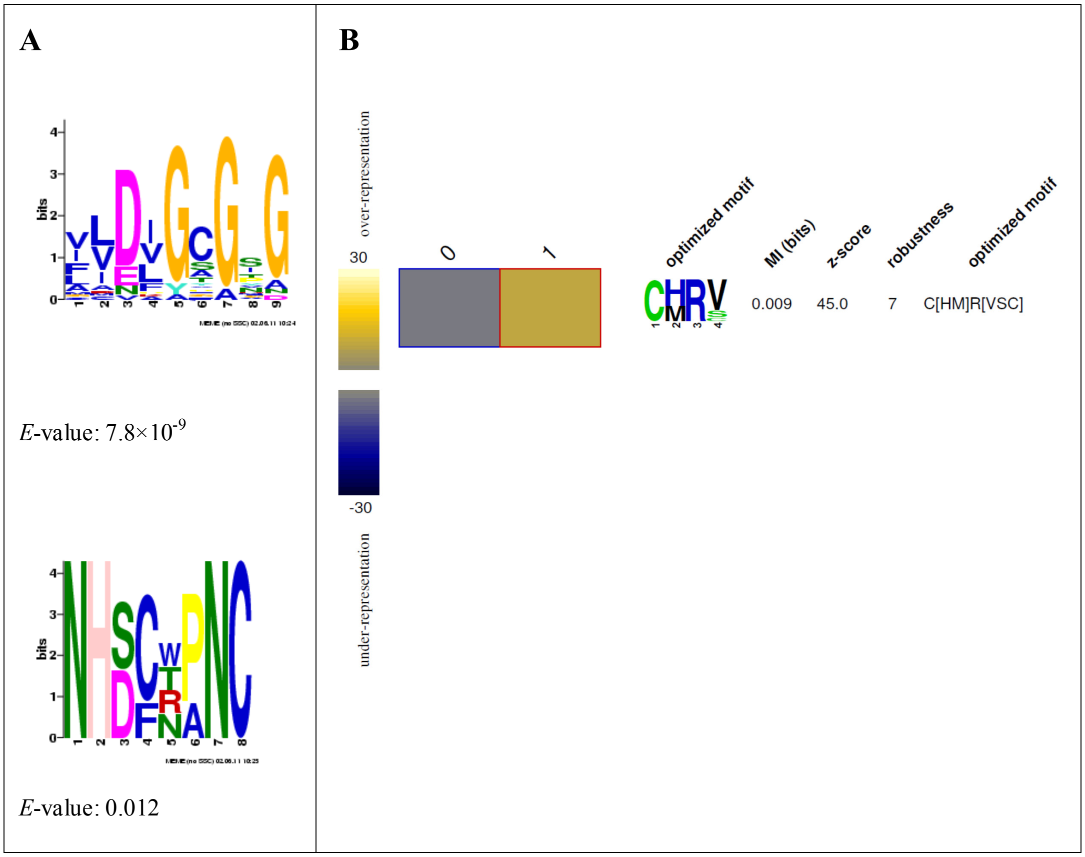
Supplementary Figure 6-2. Evaluating the performance of a classifier at 0.8 precision – For each classifier (corresponding to a particular molecular function), a *p*-value is calculated, representing the significance of the classifier sensitivity at 0.8 precision. First, the cutoff point that results in 0.8 precision is found. Then, the number of true positives (TP) and false positives (FP) that are above or below this cutoff point are determined, resulting in a contingency table as shown in this figure. The overlap of the set of true positives and the set of proteins that are above the cutoff is evaluated using a hypergeometric test, as formulated here. This figure shows the example of inorganic cation transmembrane transporter activity (GO:0022890).



Supplementary Figure 6-3. Size distribution of PROSITE patterns – PDF stands for probability density function.



Supplementary Figure 6-4. Motifs identified by MEME and FIRE-pro in adenyl ribonucleotide binding proteins – MEME identified one motif with an *E*-value smaller than 1.0 **(A)**, while FIRE-pro identified several highly degenerate short motifs **(B)**. The motif that is discovered by MEME as well as the third motif discovered by FIRE-pro resemble the ATP-binding Walker A motif. In panel **(B)**, cluster 0 represents proteins that do not have adenyl ribonucleotide binding activity, and cluster 1 represents adenyl ribonucleotide binding proteins. For more information on how to interpret this figure, see Ref. [238]. For MEME, we used the exact same set of adenyl ribonucleotide binding proteins used by HyperMotif after removal of homologs for motif discovery in the zero-or-one-per-sequence (zoops) mode. For FIRE-pro, we used this set as cluster 1, and the exact same set of negative gold standards that HyperMotif used for adenyl ribonucleotide binding activity (after removal of homologs) as cluster 0.



E-value: 7.8×10^{-9}

E-value: 0.012

Supplementary Figure 6-5. Motifs identified by MEME and FIRE-pro in methyltransferases – MEME identified two motifs with *E*-values smaller than 1.0 (A), while FIRE-pro identified one short motifs (B). The first motif that is discovered by MEME resembles the DXGCG motif discovered by HyperMotif, which directly interacts with the methyl donor substrate of methyltransfersases.

7 An overview of methods for sequence-based functional annotation

In chapters 3-6, we presented several methods for homology-independent gene function prediction, including methods for prediction of biological processes as well as a method for prediction of protein molecular functions. Before we move on to sequence-independent methods for function prediction, we may take a look back at the sequence-based methods that we described. This chapter, which was published as an opinion article in Trends in Parasitology in 2010 [4], summarizes the most prominent methods for prediction of pathways and biological processes, and conceptually compares them to our novel methods in the context of annotation of trypanosomatid genomes. The critical evaluation of the current state of annotation of parasitic genomes that is presented in this chapter 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.

7.1 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* [25, 26, 28, 29, 240], 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 about 9100 predicted and validated genes in *T. brucei*, about 4900 have no reliable homologs in the sequenced genomes of non-trypanosomatid organisms (BLAST-P, *E*-value $\leq 1\times10^{-6}$). Not all the remaining ~4200 genes can also be assigned a function, because some only have homologs that are uncharacterized too. In fact, about 35% of these conserved genes are annotated just as ‘hypothetical’. Currently, only about 3400 *T. brucei* genes have any annotation other than hypothetical (Figure 7-1).

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 hypothetical genes can be readily assigned to biological functions.

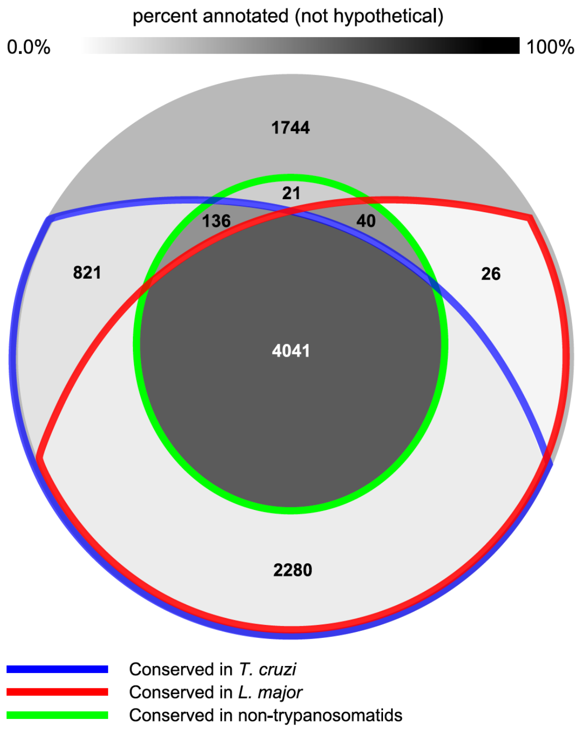


Figure 7-1. Only a small fraction of trypanosomatid genes currently have functional annotations. In this figure, *T. brucei* proteins are compared to *T. cruzi* and *L. major* proteins as well as the proteins of all other organisms with available genome sequences (blastp, E-value ≤ 1×10^{-6}). The fraction 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. Of about 6300 genes shared among *T. brucei, L. major* and *T. cruzi*, about 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) owes to 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).

7.2 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 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. [241]), (ii) the clustering of genes according to expression patterns [242] (genes with similar expression patterns have related functions [74, 243, 244]), 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 of them alone, but interaction networks claim the major share, contributing to about 85% of predictions [77]. 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 [245]. Combination of these methods has proved powerful for computational modeling of interaction networks and functional linkages [81, 246, 247]. However, many of the prominent current methods rely on the presence of homologs in other species [82, 85, 248-251], limiting their application to only a subset of genes for use in trypanosomatids.

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

A recent method, called PIC (Probabilistic-Interactome using Codon usage, Ref. [252]), 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* [252], especially 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 [253], 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% (see Ref. [253]). 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 in the next section.

7.4 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 [254]. 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 [31, 32]. However, a few regulatory elements have been identified, all of which are in the 3' UTR of developmentally regulated genes [255-278]. Some hints suggest that elements in regions other than 3' UTRs may also play roles in developmental regulation of expression [266], but none has yet been identified.

In a recent study [3], 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 potentially regulatory elements [3]. Regulatory motifs within 3' and 5' UTRs of functionally related genes were predicted using FIRE, a method that had been previously designed and applied successfully for finding informative regulatory elements [279]. 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 [3]. 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.

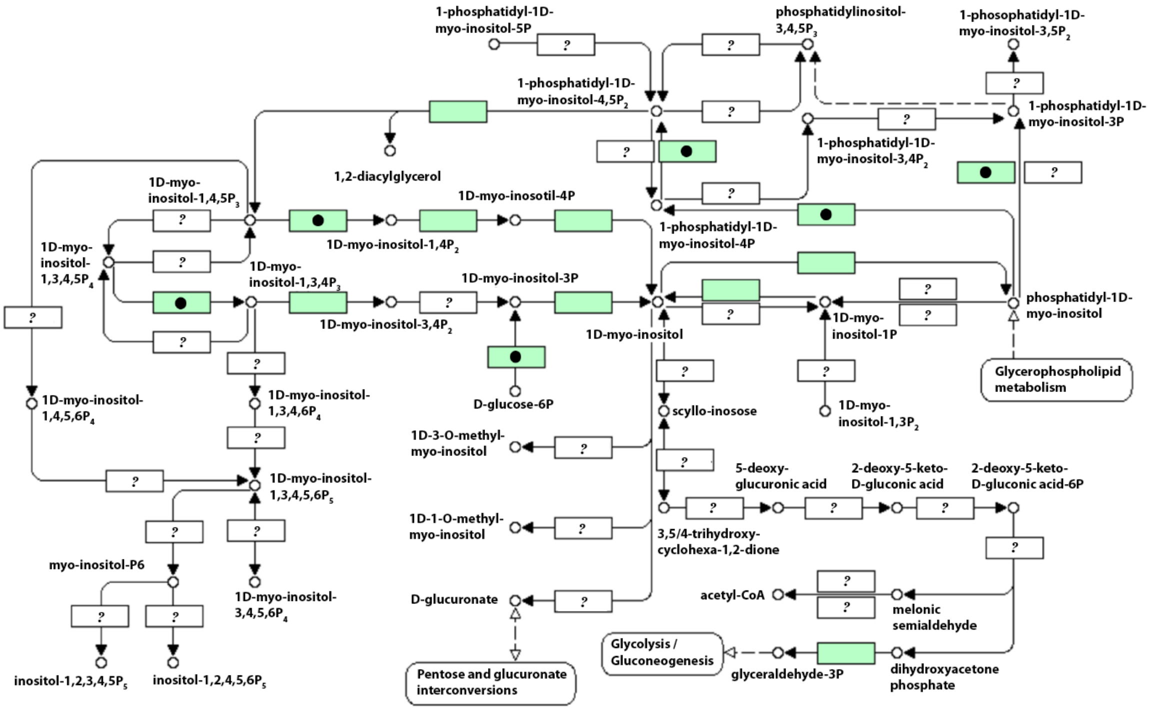

1-phosphatidyl-1D-myo-inositol-5P
phosphatidylinositol-3,4,5P$_3$
1-phosophatidyl-1D-myo-inositol-3,5P$_2$
1-phosphatidyl-1D-myo-inositol-4,5P$_2$
1-phosphatidyl-1D-myo-inositol-3P
1-phosphatidyl-1D-myo-inositol-3,4P$_2$
1,2-diacylglycerol
1D-myo-inositol-1,4,5P$_3$
1D-myo-inosotil-4P
1-phosphatidyl-1D-myo-inositol-4P
1D-myo-inositol-1,4P$_2$
1D-myo-inositol-1,3,4,5P$_4$
1D-myo-inositol-1,3,4P$_3$
1D-myo-inositol-3P
1D-myo-inositol-3,4P$_2$
1D-myo-inositol
1D-myo-inositol-1P
phosphatidyl-1D-myo-inositol
Glycerophospholipid metabolism
D-glucose-6P
1D-myo-inositol-1,4,5,6P$_4$
1D-myo-inositol-1,3,4,6P$_4$
scyllo-inosose
1D-myo-inositol-1,3P$_2$
1D-3-O-methyl-myo-inositol
1D-myo-inositol-1,3,4,5,6P$_5$
5-deoxy-glucuronic acid
2-deoxy-5-keto-D-gluconic acid
2-deoxy-5-keto-D-gluconic acid-6P
1D-1-O-methyl-myo-inositol
3,5/4-trihydroxy-cyclohexa-1,2-dione
myo-inositol-P6
1D-myo-inositol-3,4,5,6P$_4$
D-glucuronate
acetyl-CoA
melonic semialdehyde
Glycolysis / Gluconeogenesis
glyceraldehyde-3P
dihydroxyacetone phosphate
Pentose and glucuronate interconversions
inositol-1,2,3,4,5P$_5$
inositol-1,2,4,5,6P$_5$
?

Figure 7-2. 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 whose 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. This figure is drawn based on KEGG Pathway "tbr00562".

Although these motifs have not been experimentally confirmed yet, it is shown that a naïve Bayesian network can effectively predict many gene functions in *T. brucei* using the pattern of presence or absence of these predicted regulatory motifs [3]. 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 (see the previous section and Ref. [253]) and regulatory motifs [3] could make a robust gene function predictor for this particular pathway. We found out that such a combination via a simple naïve Bayesian network can achieve up to 50% sensitivity with >60% precision in identification of genes involved in inositol phosphate metabolism (Figure 7-2). We also found that 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 (*unpublished data*).

7.5 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. [280] for a review of post-translational modifications in *Plasmodium*). Proteins with similar molecular functions may also share common peptide patterns that represent their active sites [281]. In addition, functionally linked proteins may interact with a common interacting partner via similar peptide patterns [281]. All of these premises strongly suggest that function-specific protein motifs may also be exploited for predicting protein functions. Development of a tool for discovering function-specific protein motifs with a near-zero false positive rate, 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 [164]. Genes can then be clustered based on their expression patterns, often resulting in groups of biologically related genes. Each group may have a mixture of characterized and uncharacterized genes; the functions of the latter 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.

7.6 Homology-based identification of physical interactions

Rosetta stones [82], interolog mapping [248] and phylogenetic profiling [250] 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 include trypanosomatid-specific proteins as well. These methods can be combined by several means such as naïve 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.

7.7 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 the 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 both 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.

8 Functional genome annotation by combined analysis across microarray studies of *Trypanosoma brucei*

Gene expression profiles have been widely used for identification of co-regualted genes, which are usually involved in the same biological processes and pathways. However, when we started this project, no genome-wide measurements of mRNA abundance was available for *T. brucei*, which made us focus on sequence-based methods. In year 2010, several studies were published that described expression profiling of *T. brucei* genes using oligonucleotide microarrays. Soon after these publications, we performed a combined analysis of the released data in order to identify co-expressed genes, and to examine their functional linkages in *T. brucei*. In this chapter, which was published as an article in PLoS Neglected Tropical Diseases [5], we show that functional linkages among *T. brucei* genes can be identified based on gene coexpression, leading to a powerful approach for gene function prediction. These predictions can be further improved by considering the expression profiles of orthologous genes from other trypanosomatids. Furthermore, gene expression profiles can be used to discover potential regulatory elements within 3' untran slated regions. These results suggest that although trypanosomatids do not regulate genes at transcription level, trypanosomatid genes with related functions are coregulated post-transcriptionally via modulation of mRNA stability, implying the presence of complex regulatory networks in these organisms. Our analysis highlights the demand for a thorough transcript profiling of *T. brucei* genome in parallel with other trypanosomatid genomes, which can provide a powerful means to improve their functional annotation.

8.1 Background

Trypanosoma brucei, the causative agent of human sleeping sickness, is one of the major disease-causing trypanosomatids whose genome sequences have been determined for about five years [26]. However, the functions of most of the genes of this parasite still remain unknown, mainly because of the poor similarity between their sequences and the sequences of characterized genes from other organisms. This highlights the need for employing homology-independent approaches to improve the functional annotation of *T. brucei* genome. Since co-expressed genes tend to share similar functions, belong to the same pathways, or participate in the same processes [282], the function of a gene can often be predicted based on the functions of the genes it is co-expressed with [283]. This provides a powerful homology independent method for functional annotation of a genome.

In *T. brucei*, most genes are not transcriptionally regulated [31, 32]. Instead, genes are transcribed as polycistronic mRNAs [93] that heavily depend on post-transcriptional processes for maturation and regulation. Some reports suggest that this lack of transcriptional regulation results in limited responsiveness of *T. brucei* transcriptome to altered environment and genetic background [208], thus, preventing the construction of an informative coexpression network. Nevertheless, recent studies have reported that mRNAs of *T. brucei* genes with related functions share similar sequence motifs in their untranslated regions (UTRs), suggesting that they are coregulated at post-transcriptional level via common sequence-dependent mechanisms for regulation of mRNA stability and/or translation [3].

Three recent studies have provided genome-wide expression profiles for procyclic form (PF) and bloodstream form (BF) *T. brucei* during differentiation [34, 35, 164]. Here, we demonstrate that while the data from each of these individual studies is not significantly informative about gene function, their collection can be used to construct a coexpression network that reflects the functional linkages among genes. We have used this coexpression network to predict the broad functions of several currently uncharacterized *T. brucei* genes, and have expanded our predictions by considering coexpression relationships that are conserved between *T. brucei* and *Leishmania infantum*. Finally, we show that by combining the expression data from the microarray studies of *T. brucei*, we can cluster the genes based on expression profiles and use these clusters to identify potential regulatory elements within mRNA untranslated regions.

8.2 Methods

The methods that we have used in this study are summarized in this section. The details of the methods are provided in Supplementary Methods.

8.2.1 Data sources

We used *T. brucei* mRNA expression data from three recent publications [34, 35, 164]. A set of 7488 *T. brucei* genes was shared by these three studies, each gene represented by a total of 17 expression values: four from ref. [34], eight from ref. [164] and five from ref. [35]. The functional annotations of *T. brucei* genes were obtained from KEGG pathway database [284] and TriTrypDB [285]. The sequences of 3 UTRs were extracted based on previous splice-site predictions [47]; sequences were either used completely or truncated to contain only the first 1000nt in the 5end of the 3′ UTR. For identification of conserved coexpression, we used a collection of *Leishmania infantum* gene expression profiles from three different studies [286-288]. Orthologous genes between *T. brucei* and *L. infantum* were identified based on their protein sequences, obtained from KEGG [284].

8.2.2 Construction and evaluation of a coexpression network based on T. brucei microarray studies

The coexpression values for $\sim 2.8\times10^{7}$ *T. brucei* gene pairs, measured as Pearson correlation coefficients across several experiments, were obtained using different experiment sets: (i) a set of four experiments from ref. [34], (ii) a set of eight experiments from ref. [164], (iii) a set of five experiments from ref. [35], (iv) the set of all the 17 experiments from these three studies, and (v) a selected subset from the 17 experiments; this subset was chosen so as to maximize the accuracy and coverage of predicting functional linkages, as explained in the next section. Gene pairs with correlation coefficients greater than a specified threshold were used to construct the coexpression networks. This threshold was chosen so that at least 75% of linkages in the coexpression network would represent functional linkages according to KEGG (in other words, the coexpression network would have a precision of 75%).

8.2.3 Selecting an optimum subset of microarray experiments for identification of functional linkages

Different microarray studies may present data that do not equally correlate with functional linkages; inclusion of experiments that do not reflect the functional relationships among genes may have a negative effect on the accuracy of function predictions. Furthermore, some experiments may be redundant; e.g. replicate the same biological condition or show little differences in terms of the transcriptome profile. Therefore, it is necessary to trim the dataset that is used for construction of the coexpression network in order to remove redundant and uninformative experiments. To this end, we used a heuristic algorithm for selection of the best subset. This algorithm tries to iteratively find experiments whose exclusion can actually improve the accuracy and coverage of the coexpression network. It should be noted that although these 'excluded' experiments may have a negative effect on the 'overall' accuracy of function predictions, they may provide specific information for particular pathways, as we will show in the results.

8.2.4 Gene function prediction based on the coexpression network

The functions of currently uncharacterized genes can be predicted based on their association with genes of known functions in the coexpression network. Briefly, if a particular gene is coexpressed with several genes that have a shared function, that gene is also most likely involved in the same function. We calculated a p-value for each gene-pathway pair, so that a small p-value would reflect a significant association between the gene and the pathway. Uncharacterized genes were assigned to biological pathways if their association had a p-value that corresponded to at least 80% precision, meaning that at least an estimated 80% of the predictions are correct.

8.2.5 *Identification of conserved coexpression linkages among genes*

Genes with related functions have usually conserved their coexpression through evolution. Thus, if two genes are coexpressed in more than one organism, there is a higher chance that these genes are functionally related [283]. We identified 5300 orthologs of *T. brucei* genes in the closely related organism *Leishmania infantum* based on reciprocal best BLASTP hits with e-values $<1\times10^{-6}$. The coexpression value in *L. infantum* was calculated for gene pairs based on a collection of previously reported data from three different studies [286-288]. Each pair of conserved genes could then be assigned two values: their Pearson correlation coefficient based on *T. brucei* microarray data, and their Pearson correlation coefficient based on *L. infantum* microarray data. Two genes have a conserved coexpression relationship if both of these values are greater than specified cutoffs (different cutoffs can be used for each organism). The cutoffs were chosen so that the conserved coexpression network would have maximum coverage of *T. brucei* proteins with a precision of at least 50%.

8.2.6 *Identification of potential regulatory motifs in UTRs*

We used a previously reported regulatory element discovery method, FIRE, which has been shown to have a close-to-zero false discovery rate and provides a wealth of information about each of the discovered motifs [168]. *T. brucei* genes were clustered based on the data of the three microarray studies [73], and the gene clusters along with either complete or truncated **3**UTR sequences were submitted to FIRE with default parameters. We only discuss the results of running FIRE on truncated sequences in this paper; the complete set of results can be found at http://webpages.mcgill.ca/staff/Group2/rsalav/web/Suppl/20100109/index.htm.

8.3 Results and Discussion

8.3.1 *A coexpression network of T. brucei genes*

We calculated the pairwise correlation coefficients of mRNA expression profiles for ~2.8×10^{7} gene pairs in each of the three *T. brucei* microarray datasets as well as in a combined dataset. As Figure 8-1A shows, if each dataset is considered separately, only a minor enrichment of functionally associated genes can be observed at high correlation coefficients. Nonetheless, when the three datasets are merged, the enrichment ratio of functional linkage between coexpressed genes increases drastically, reaching as high as ~20 for gene pairs with correlation coefficients >0.90. This can be further improved by objectively selecting the set of experiments that are used for calculating correlation coefficients: removing all the experiments from ref. [34], three out of eight experiments from ref. [164] and one of the five experiments from ref. [35] could increase the enrichment of functionally associated genes up to three-fold for gene pairs with correlation coefficients >0.95 (Figure 8-1A). This also significantly improved the accuracy of predicting functionally associated gene pairs (Figure 8-1B). However, as we will show later, while the trimmed dataset is generally more successful in identification of functionally associated genes, the non-trimmed dataset can better identify genes of particular functions, such as oxidative phosphorylation.

Using the combined microarray datasets, we constructed two coexpression networks of *T. brucei* each with an estimated precision of 75% (Figure 8-1C and D). We call the network that is obtained from all microarray experiments $\text{CoExp}^{1}_{\text{Tbr}}$ and the network that is obtained from the selected subset of experiments $\text{CoExp}^{2}_{\text{Tbr}}$. These networks encompass 1280 and 10247 connections among 799 and 4148 *T. brucei* genes, respectively. Most of these genes have no known function (49% in $\text{CoExp}^{1}_{\text{Tbr}}$ and 59% in $\text{CoExp}^{2}_{\text{Tbr}}$ are annotated as hypothetical proteins).

The $\text{CoExp}^{1}_{\text{Tbr}}$ network consists of two main clusters, one with a large number of bloodstream form (BF)-specific genes and one with mostly procyclic form (PF)-specific genes. Some protein complexes and functional modules can be readily distinguished in the sub-network that has most of the PF-specific proteins, as shown in Figure 8-1C. This modularity of the network should allow us to predict the functions of currently uncharacterized genes. For example, Tb927.10.4880 (formerly identified as Tb10.70.2320), which is currently annotated as “hypothetical conserved”, is located within a complex that corresponds to cytochrome c oxidase. This is congruent with the recent reports showing that this protein co-purifies with cytochrome c oxidase complex [289].

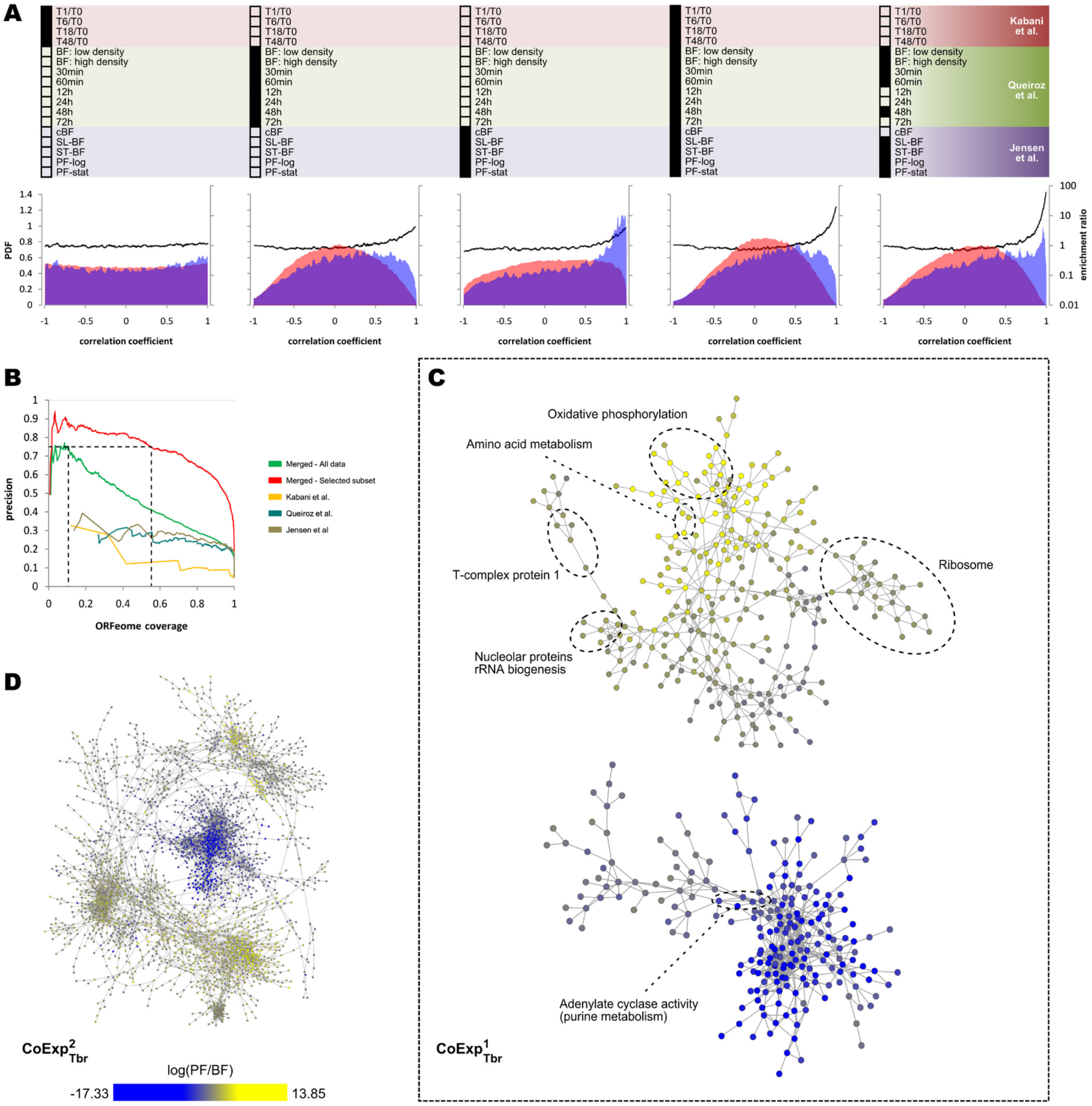


Figure 8-1. Integration of microarray data for identification of functional linkages among genes – (A) The correlation coefficients between genes were calculated for each *T. brucei* dataset separately, for the combination of the three datasets, and for a selected subset of the experiments. The probability density function (PDF) of correlation coefficients among functionally associated and non-associated genes is shown by blue and red, respectively. It can be seen that the data from the work by Kabani et al. [34] are poorly correlated with functional linkages. This is while the other two datasets from Queiroz et al. and Jensen et al. [35, 164] can discriminate functionally linked gene pairs based on the higher correlations of their

expression profiles. Consequently, the procedure that we used for selection of the best subset of the experiments automatically excluded the data from Kabani et al. [34], while retaining most of the experiments from the other two datasets (the right panel). The enrichment of functional linkages at a given correlation coefficient, shown by the thick black line, was calculated by dividing the values of the two PDFs. **(B)** Precision (positive predictive value, PPV) vs. ORFeome coverage for prediction of functional linkages based on coexpression is shown in this graph. ORFeome coverage is defined as the fraction of ORFs (open reading frames) with associated expression profiles that are coexpressed with at least one other ORF. By decreasing the threshold for identification of coexpressed pairs, more ORFs are included in the network, but the fraction of coexpression relationships that reflect functional linkages (i.e. precision) decreases. At a precision of 0.75, $\mathrm{CoExp}^1_{\mathrm{Tbr}}$ and $\mathrm{CoExp}^2_{\mathrm{Tbr}}$ include 10.7% and 55.4% of *T. brucei* ORFeome, respectively. The correlation coefficient cutoff for $\mathrm{CoExp}^1_{\mathrm{Tbr}}$ is 0.94 and for $\mathrm{CoExp}^2_{\mathrm{Tbr}}$ is 0.957. **(C)** In $\mathrm{CoExp}^1_{\mathrm{Tbr}}$, functionally related genes cluster together. A global view of $\mathrm{CoExp}^2_{\mathrm{Tbr}}$ is also provided in panel **(D)**. Stage-specific expressions are shown by node colors, with yellow for PF-specific and blue for BF-specific proteins. These two networks can be downloaded at http://webpages.mcgill.ca/staff/Group2/rsalav/web/Suppl/20100109/index.htm.

Based on visual inspection, the sub-network with BF-specific proteins has notably less modularity compared to the PF-enriched sub-network. Although several functions are enriched among BF-specific genes (Supplementary Figure 8-1), they are not represented adequately in this coexpression network due to its low coverage. However, as expected from the higher coverage of $\mathrm{CoExp}^2_{\mathrm{Tbr}}$, this network contains more BF-specific genes. It can be anticipated that upon the availability of more microarray data, both the coverage and the precision of the coexpression network will be even further improved and, consequently, a more modular and thorough coexpression network will emerge. Nonetheless, the current networks can be used to predict the functions of many currently hypothetical *T. brucei* genes, as explained in the next section.

8.3.2 Pathways can be predicted based on coexpression networks of *T. brucei*

As Figure 8-1C shows, genes of different functions are clustered together in the coexpression network of *T. brucei*. We used this functional relatedness of coexpressed genes to predict functions of uncharacterized genes within the obtained networks. As shown in Figure 8-2, each of the $\text{CoExp}^1_{\text{Tbr}}$ and $\text{CoExp}^2_{\text{Tbr}}$ networks are more successful in finding new genes for different pathways: $\text{CoExp}^1_{\text{Tbr}}$ can successfully assign new genes to ribosome, oxidative phosphorylation and purine metabolism, while $\text{CoExp}^2_{\text{Tbr}}$ can identify genes that are involved in ribosome, glycolysis, inositol phosphate metabolism and phosphatydilinositol signaling system (the genes involved in the latter two pathways considerably overlap, according to KEGG pathway annotations).

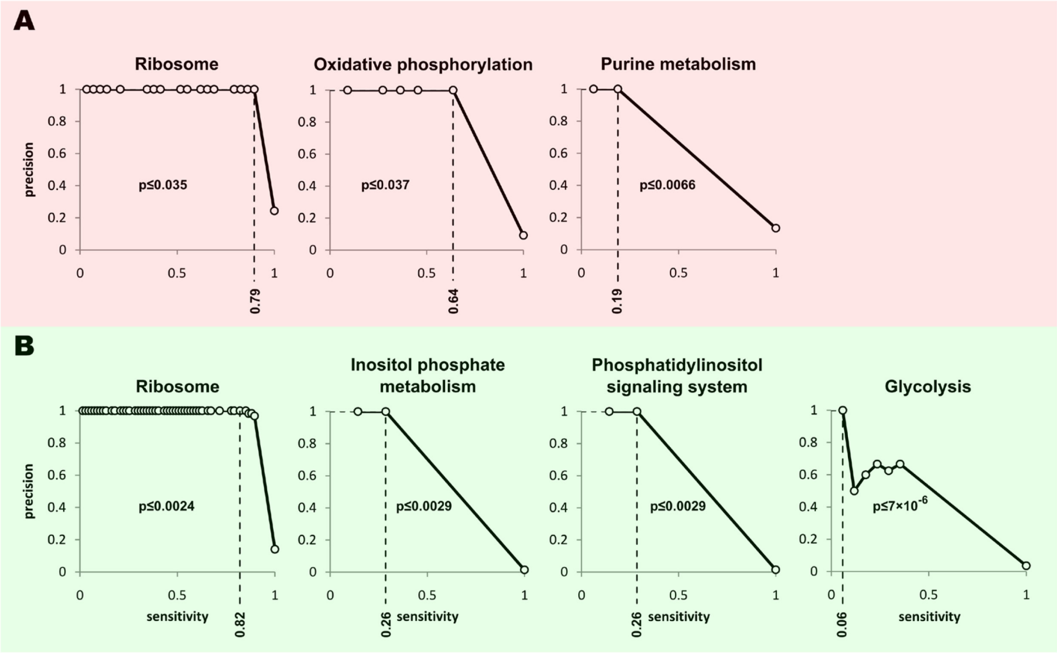


Figure 8-2. Function prediction based on *T. brucei* coexpression networks – Precision-recall curve for each function is plotted separately. Recall or sensitivity for a particular pathway is defined as the fraction of genes of that pathway within the coexpression network whose function is correctly predicted. Precision indicates the fraction of the predictions that are correct. The $\text{CoExp}^1_{\text{Tbr}}$ network can successfully predict ribosome, oxidative phosphorylation, and purine metabolism genes **(A)**, while $\text{CoExp}^2_{\text{Tbr}}$ is best at predicting ribosome, inositol phosphate metabolism, phosphatidylinositol signalling system, and glycolysis

genes **(B)**. The p-value thresholds were chosen to be at most 0.05 and result in a precision of at least 0.8. See **Supplementary Table 8-1** and **Supplementary Table 8-2** for complete sets of predictions.

We found that many of the genes whose functions are predicted based on our analysis, although having no annotation in KEGG pathway database, are already annotated in TriTrypDB [285]. These annotations are considerably congruent with our predictions, highlighting the reliability of our approach in predicting gene functions. Examples include several cytochrome c oxidase subunits that are correctly assigned to oxidative phosphorylation and many 40S and 60S ribosomal proteins that are correctly assigned to ribosome. While this provides a proof of concept for the method, it also underpins the limitations of KEGG pathway database as the gold standard for construction of the functional linkage network and subsequent function prediction. For example, the gene Tb927.10.4880, which we mentioned in the previous section, cannot be assigned to any function using KEGG pathway information, since none of its neighbors in the coexpression networks are annotated in KEGG. However, if we manually add the known cytochrome c oxidase subunits of *T. brucei* to the oxidative phosphorylation pathway in the gold standard set, our approach can successfully predict that Tb927.10.4880 is involved in oxidative phosphorylation ($p<0.001$).

Nonetheless, based on the coexpression networks, we can readily predict the likely pathways and biological processes for many of the currently hypothetical proteins. Some of these predictions are also corroborated with available literature. For example, Tb927.10.9830 (formerly identified as Tb10.6k15.0480), which, based on $\text{CoExp}^1_{\text{Tbr}}$, is predicted to be involved in oxidative phosphorylation, has been previously reported to be associated with ATP synthase complex [290]. Tb927.4.4020 and Tb927.10.7090 (formerly known as Tb10.6k15.3640) which are coexpressed with purine metabolism genes have several copies of putative regulatory elements that have been previously reported as purine metabolism-specific 3′ UTR motifs [3]. Also, Tb927.6.2330, which, based on $\text{CoExp}^2_{\text{Tbr}}$, is predicted to be associated with ribosome, has an RGG domain which has been shown to interact with several ribosomal proteins [291]. The complete list of our predictions based on $\text{CoExp}^1_{\text{Tbr}}$ and $\text{CoExp}^2_{\text{Tbr}}$ along with literature information that either support or oppose these predictions can be found in Supplementary Table 8-1 and Supplementary Table 8-2. The distribution of these genes in the coexpression networks is shown in Supplementary Figure 8-2.

We have also used the coexpression networks $\mathrm{CoExp}^{1}_{\mathrm{Tbr}}$ and $\mathrm{CoExp}^{2}_{\mathrm{Tbr}}$ to predict the likely biological processes, molecular functions, and cellular compartments of *T. brucei* genes based on GO annotations of TriTrypDB (Supplementary Table 8-3 to Supplementary Table 8-8). The analysis of GO annotations complements the KEGG dataset by expanding the predictions of metabolic pathways and also by providing predictions for other categories. For example, we were able to predict novel genes that are potentially involved in antigenic variation, protein folding, and microtubule-based movement. Also, this analysis showed that many proteins within the same cellular compartments are coexpressed, which is not surprising as cellular compartmentalization loosely reflects functional compartmentalization of proteins. This allowed us to predict the likely localization of many proteins; most notably we were able to find potential membrane proteins, intracellular proteins, and proteins associated with dynein complex (Supplementary Table 8-5 and Supplementary Table 8-8).

8.3.3 Conserved coexpression: a closer look

Conservation of coexpression is a much stronger indicative of functional linkages among genes, compared to coexpression in a single organism [283, 292]. Thus, we searched for coexpression associations that were conserved between *T. brucei* and its close relative, *L. infantum*. As Figure 8-3A shows, in the subset of genes whose orthology between *T. brucei* and *L. infantum* can be unambiguously established, gene pairs that are coexpressed in both *T. brucei* and *L. infantum* are considerably enriched with functional linkages. This property can be used for a more accurate prediction of functional linkages, as shown in Figure 8-3B: while neither the microarray data of *T. brucei* nor those of *L. infantum* alone can reach a precision higher than 40% for identification of functional linkages among the conserved subset of genes, their combination can yield a wide range of precision and sensitivity values (note that the lower precision of *T. brucei*-only data compared to $\mathrm{CoExp}^{1}_{\mathrm{Tbr}}$ reflects the absence of most of ribosomal proteins from the subset of genes with unambiguous orthologs; see Supplementary Figure 8-3).

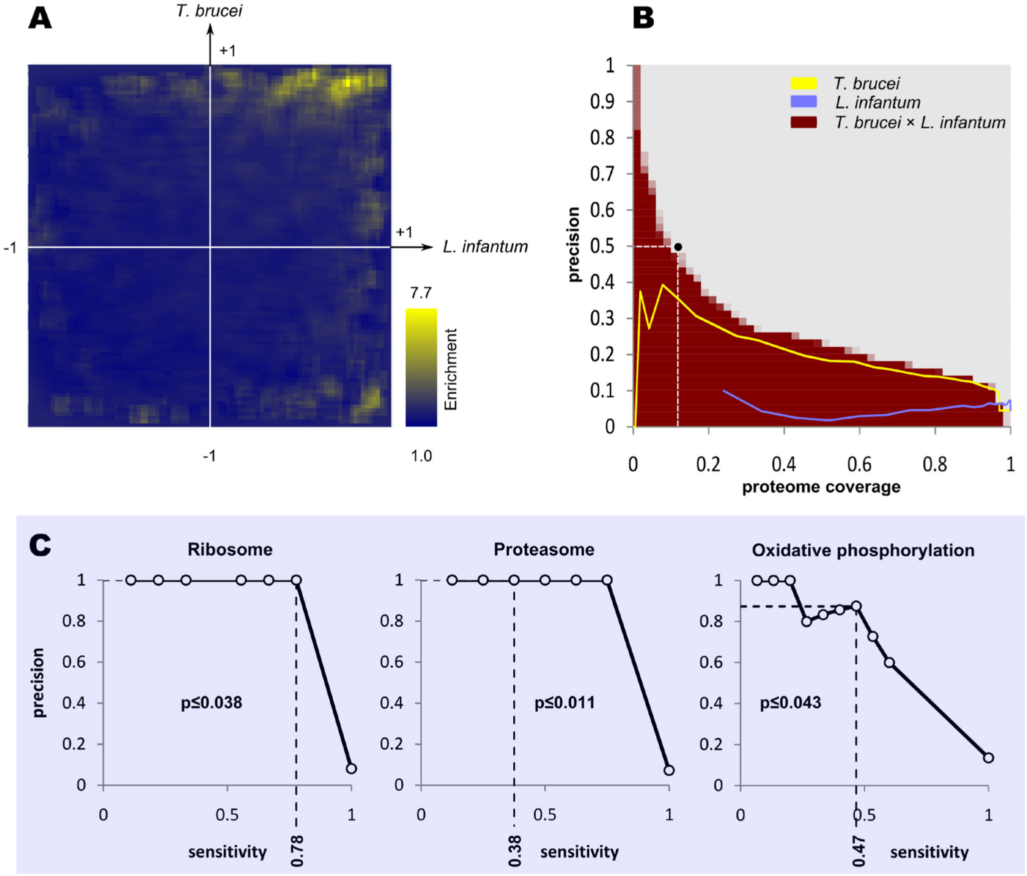


Figure 8-3. Prediction of functional linkages based on conservation of coexpression – (A) Gene pairs that are functionally related are coexpressed in both *T. brucei* and *L. infantum*. Therefore, an enrichment of functional linkages can be observed where correlation coefficients are high for both *T. brucei* and *L. infantum* (the *x*-axis represents the correlation coefficients of gene pairs in *L. infantum*, while the *y*-axis represents that correlation coefficients in *T. brucei*). **(B)** By considering the conservation of coexpression between *T. brucei* and *L.infantum* (red), we can more accurately predict functional linkages, compared to predictions that are based solely on *T. brucei* data (yellow) or *L. infantum* (light blue). About 50% of gene pairs whose expression profiles have correlation coefficients greater than 0.89 in *T. brucei* and 0.56 in *L.infantum* are estimated to be functionally related (black circle). These gene pairs cover ~11.9% of all *T. brucei* genes with unambiguous *L. infantum* orthologs. The resultant conserved coexpression network can be downloaded at http://webpages.mcgill.ca/staff/Group2/rsalav/web/Suppl/20100109/index.htm. **(C)**

Ribosome, proteasome and oxidative phosphrylation genes can be identified based on the conserved coexpression network. See **Supplementary Table 8-9** for complete set of predictions.

We chose our criteria for identification of conserved coexpression relationships so that at least 50% of these relationships reflect functional linkages among genes. This resulted in a conserved coexpression network with 1110 associations among 632 *T. brucei* genes whose orthologs in *L. infantum* could be unambiguously identified. Based on this network, many new genes could be mapped to KEGG pathways (Figure 8-3 3C, Supplementary Figure 8-2). This conserved coexpression network was particularly successful in assigning currently uncharacterized genes to oxidative phosphorylation (Supplementary Table 8-9). For example, from 17 hypothetical conserved genes that based on this network were predicted to be involved in oxidative phosphorylation, seven genes have been previously identified as potential associated partners or subunits of ATP synthase complex [290]; five others have been reported as mitochondrial proteins, one of which is specifically identified as a mitochondrial membrane protein [293, 294]; and three proteins have a potential regulatory element in their transcript that is also found in the transcripts of many cytochrome c oxidase subunits [61]. This conserved coexpression network could also be used for predicting the likely GO associations of a few *T. brucei* genes (Supplementary Table 8-10).

These results suggest that functionally related genes are coregulated at mRNA level, most probably through post-transcriptional processes, in different trypanosomatids including both *Trypanosoma* and *Leishmania* genera. Furthermore, this analysis highlights the parallel expression profiling of trypanosomatids as a promising approach that can significantly enhance the functional annotation of all trypanosomatid genomes, including *T. brucei*.

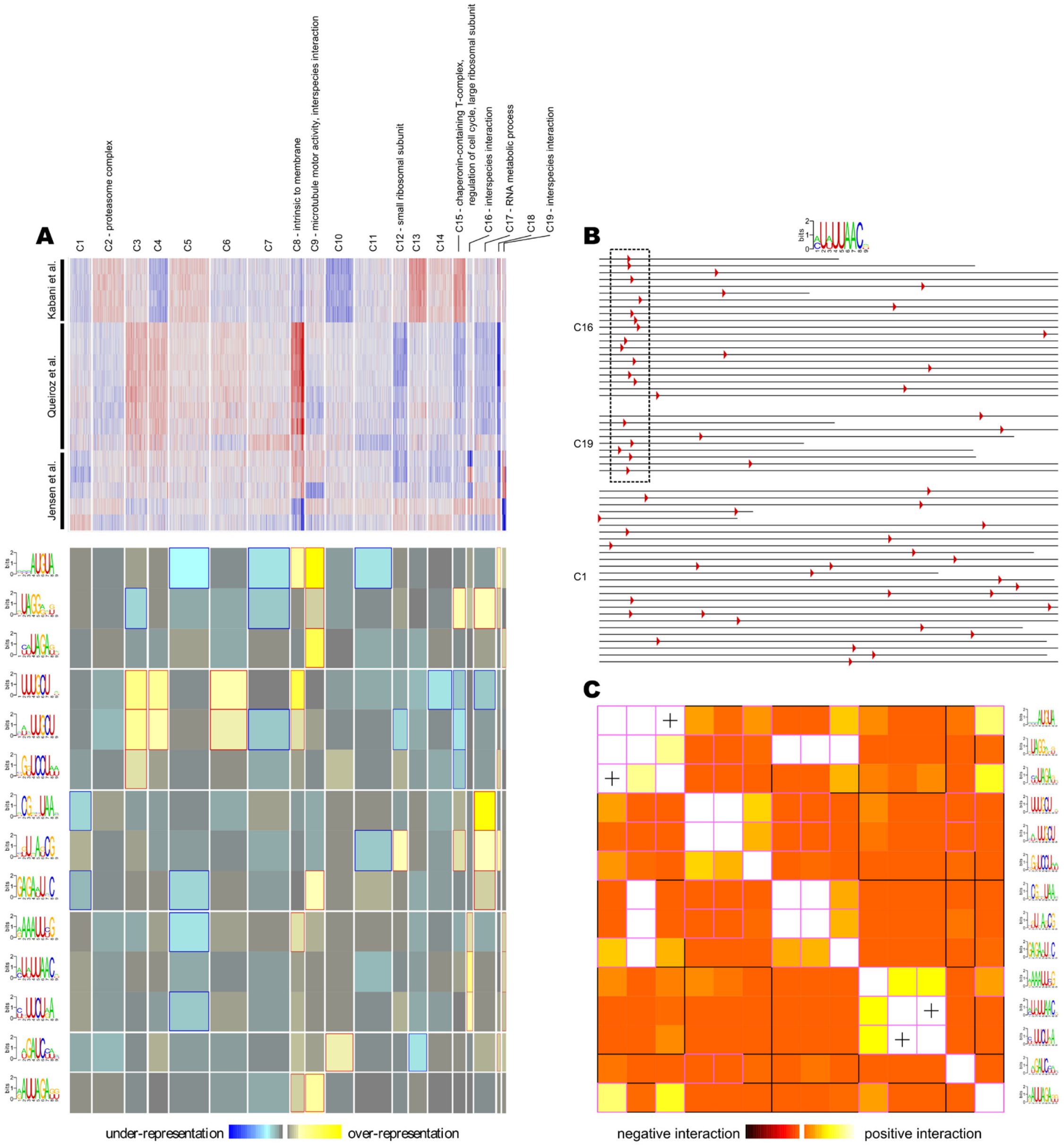


Figure 8-4. Finding potential regulatory elements based on a combined microarray dataset – (A) *T. brucei* genes were grouped into 19 clusters based on their expression profiles (top panel; red: high expression, blue: low expression). FIRE [168] was used to find potential regulatory motifs in the 3′ UTRs (lower panel; yellow and blue represent over-representation and under-representation of a motif within a cluster, respectively). **(B)** The motif [AC]U[AU]UUAAC occurs preferably between nucleotides 40 and 100 downstream of the stop codon in clusters 16 and 19, while its position is random in other clusters, such as cluster 1. Interestingly, both clusters 16 and 19 are enriched with genes involved in interspecies

interaction (mostly surface antigens). **(C)** Some motif pairs co-occur in the 3′ UTRs. In this symmetric heat map, each row and each column corresponds to a predicted motif. Light color indicates that the presence of a motif in a 3′ UTR implies the presence of another motif within the same UTR. Significant spatial co-localization between pairs of motifs is shown by "+". The full set of results along with additional analyses can be found at http://webpages.mcgill.ca/staff/Group2/rsalav/web/Suppl/20100109/index.htm.

8.3.4 Cis-regulatory element discovery based on clusters of coexpressed genes

Having a collection of microarray datasets, we can study the underlying mechanisms of gene regulation in *T. brucei*. To this end, we first clustered the *T. brucei* genes based on their expression profiles into 19 distinct coexpression groups. The expression patterns within each of these clusters were consistent, and different clusters had unique signatures distinguishing them from each other (Figure 8-4A). Eight out of the 19 clusters significantly overlapped with at least one Gene Ontology (GO) category, including biological processes, molecular functions, and cellular compartments ($p<0.05$ with Bonferroni correction for multiple comparisons; all GO terms at all levels were analyzed). We next used FIRE [168] to find potential regulatory motifs in 3UTRs across these clusters. FIRE was able to find 14 statistically significant RNA motifs, each over-represented in different gene clusters (Figure 8-4A). Interestingly, some of these motifs showed a position bias in the clusters in which they were over-represented. For example, the motif [AC]U[AU]UUAAC, which is over-represented among genes that are involved in the interaction of parasite with host, occurs mostly between the 40th and 100th nucleotide after the stop codon of these genes, while showing no position preference in UTRs of genes of other clusters (Figure 8-4B). Furthermore, many of the found motifs seem to co-occur within the same UTR (Figure 8-4C). This suggests that they may represent the most conserved parts of a larger, probably structural, RNA motif. This is especially the case for predicted motifs that not only co-occur with each other, but also co-localize at the same part of the 3′ UTR.

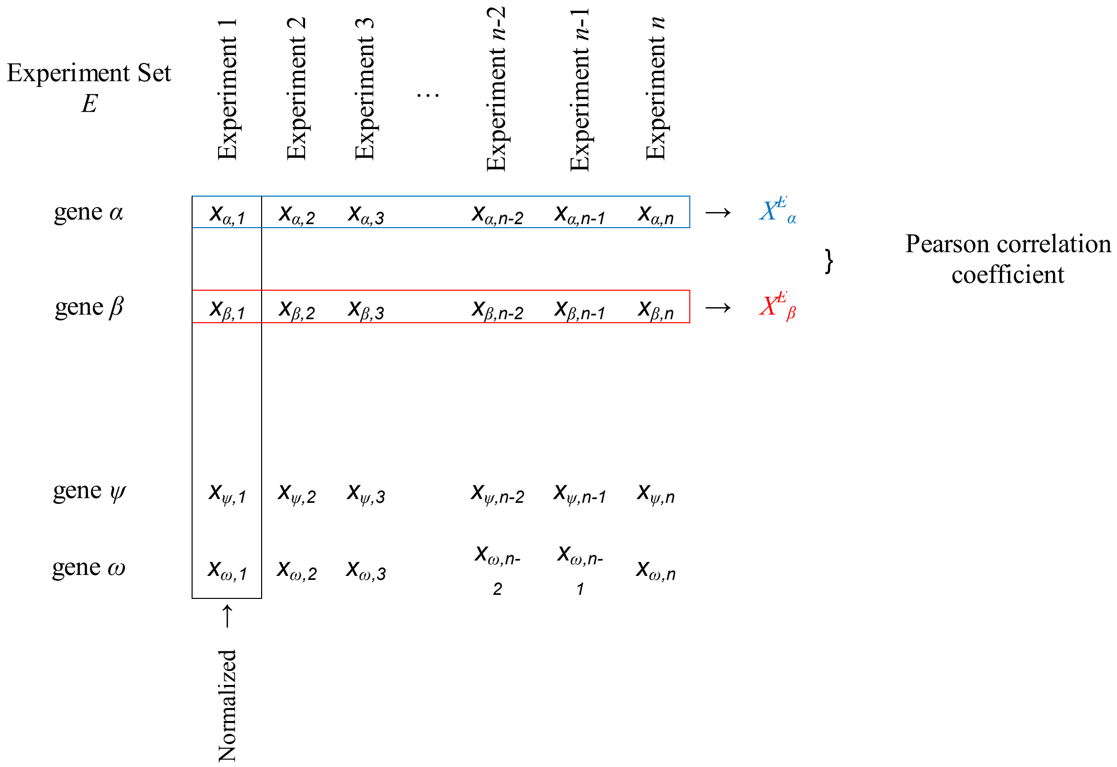
8.3.5 Concluding remarks

The analysis presented in this paper highlights whole-genome transcript profiling as a powerful tool for identification of functional and regulatory modules in *T. brucei*. A comprehensive and high-resolution analysis, however, needs tens to hundreds of different microarray experiments in order to capture the nuances between gene expression patterns of different modules. These experiments should encompass a variety of environmental and genetic conditions, including different stress-inducing culture media and various knockdown/knockout cells. Nonetheless, recent studies suggest that once a large collection of microarray data is available, regulatory and functional modules may be identified even in the absence of such environmental and genetic variations [295]. It should also be noticed that parallel transcript profiling of related organisms can potentially provide more information than excessively thorough transcript profiling of a single organism.

8.4 Supplementary Methods

8.4.1 Construction of the coexpression networks

Expression values in each microarray experiment, expressed as log ratios of the signal from experimental cDNA to the signal from reference cDNA, were normalized to have a mean of 0.0 and a standard deviation of 1.0. This was done by calculating the average (μ) and standard deviation (σ) for each experiment, and transforming each value by subtracting the average and dividing by the standard deviation: $x=(x'-\mu)/\sigma$, where x' is the original value and x is the transformed (normalized) value. Given two genes α and β from S (S is the set of all genes with associated expression profiles) and their normalized expression values across different experiments of the experiment set E, the coexpression value of α and β can be calculated as the Pearson correlation coefficient of X^{E}_{α} and X^{E}_{β}, where X^{E}_{α} represents the measurements for α in the set E, and X^{E}_{β} represents the measurements for β in the set E, as shown in the figure below:



The coexpression network G^E_θ is the set of all gene pairs whose Pearson correlation coefficients, according to the experiment set E, are at least θ:

$G^E_\theta=\{(i,j)|\rho(X^E_i,X^E_j)\geq\theta\}$,

where ρ is the Pearson correlation coefficient function. The set of nodes in the coexpression network G^E_θ is denoted as N^E_θ:

$N^E_\theta=\{i|\exists j:\rho(X^E_i,X^E_j)\geq\theta\}$

$|N^E_\theta|$ therefore represents the number of nodes in the network G^E_θ. The coverage of the network is defined as:

$f^E_\theta=|N^E_\theta|/|S|$

Thus, f^E_θ indicates what fraction of all genes the network G^E_θ represents. A higher coverage implies that the network can potentially be used for prediction of functions for a larger fraction of *T. brucei* genes with available expression profiles.

The precision of a network in finding functional interactions is calculated by comparing the network to gold standard positive and negative sets. The gold standard positive set I consists of all gene pairs that share at least one function according to KEGG pathway database:

$I=\{(i,j)|F_i\cap F_j\neq\emptyset\}$,

where F_i and F_j represent the set of functions for genes i and j according to KEGG. The gold standard negative set I' includes all gene pairs that do not share any function, given that each gene has at least one annotation in KEGG pathway database:

$I'=\{(i,j)|F_i\cap F_j=\emptyset,F_i\neq\emptyset,F_j\neq\emptyset\}$

The term “tbr01100” (Metabolic pathways) was ignored in all analyses.

The limitations and incompleteness of both I and I' need to be noted: not all *T. brucei* genes with known functions are represented in KEGG; therefore, I is far from complete. Furthermore, the annotations for genes that are present in KEGG may not be complete, meaning that two genes may actually share a pathway, but this information is missing from KEGG; therefore, I' may contain some gene pairs that should actually belong to I but are mistakenly assumed as negatives.

The positive predictive value (PPV, also referred to as precision) of the network G^E_θ is defined as:

$p^E_\theta=|G^E_\theta\cap I|/(|G^E_\theta\cap I|+|G^E_\theta\cap I'|)$

Therefore, p^E_θ estimates the fraction of gene pairs in G^E_θ that are functionally related. We used the area under the curve (AUC) for $p^E_\theta(\theta)$ vs. $f^E_\theta(\theta)$ as an estimate of how well the experiment set E can reflect the functional linkages among genes. This AUC is here referred to as A^E.

In this study, we used different experiment sets: E_K which is the set of four experiments from ref. [34], E_Q which is the set of eight experiments from ref. [164], E_J which is the set of five experiments from ref. [35], $E_{KQJ}= E_K+E_Q+E_J$, and $\ddot{E}\subseteq E_{KQJ}$. $\ddot{E}$ is chosen so as to result in the maximum A^E:

$A^{\ddot{E}}\geq A^E\ \forall E\subseteq E_{KQJ}$

Since all subsets of E_{KQJ} could not be tested due to computational limitation, we used a heuristic approach to find $\ddot{E}$. A pseudocode for this approach is shown below:

```
1. Set $\ddot{E}=E_{KQJ}$
2.     Create the list $L=\{E'|E'\subseteq E_{KQJ}, |E'|=|\ddot{E}|-1 \vee |E'|=|\ddot{E}|+1\}$
3.     Find the $E$ in $L$ that has the maximum $A^E$
4.     If $A^E>A^{\ddot{E}}$ then set $\ddot{E}=E$ and go to step 2
5. Report $\ddot{E}$
```

Using each of the experiment sets E_{KQJ} and $\ddot{E}$, we defined a coexpression network by selecting the minimum value for cutoff θ that could result in $p_\theta\geq 0.75$ (i.e. precision of at least 75%). The selected value of θ for E_{KQJ} was 0.94 and for $\ddot{E}$ was 0.957. The resulting networks are referred to in the paper as $\text{CoExp}^1_{\text{Tbr}}$ and $\text{CoExp}^2_{\text{Tbr}}$, respectively.

8.4.2 Identification of conserved coexpression linkages among genes

We identified 5300 orthologs of *T. brucei* genes in the closely related organism *Leishmania infantum* based on reciprocal best BLAST-P hits with e-values $<1\times10^{-6}$. The set of *T. brucei* genes whose *L. infantum* orthologs could be unambiguously identified is referred to as S'. The experiment set E' for *L. infantum* was obtained from three different studies [286-288]. The conserved coexpression network $G^{E,E'}_{\theta,\theta'}$ is the set of all gene pairs that are coexpressed according to both experiment sets $E=E_{KQJ}$ (for *T. brucei*) and E' (for *L. infantum*):

$G^{E,E'}_{\theta,\theta'}=\{(i,j)|i\in S',j\in S',\rho(X^E_i,X^E_j)\geq\theta,\ \rho(X^{E'}_{i'},X^{E'}_{j'})\geq\theta'\}$, where ρ is the Pearson correlation coefficient function, i' is the ortholog of i in *L. infantum* and j' is the ortholog of j in *L. infantum*. To identify the best θ and θvalues, we tried all pairs of values so that $\theta \in \{-1,-0.99,-0.98,\ldots,0.98,0.99,1\}$ and $\theta' \in \{-1,-0.99,-0.98,\ldots,0.98,0.99,1\}$. The pair of values that resulted in the maximum coverage of S' and a precision of at least 0.50 was chosen.

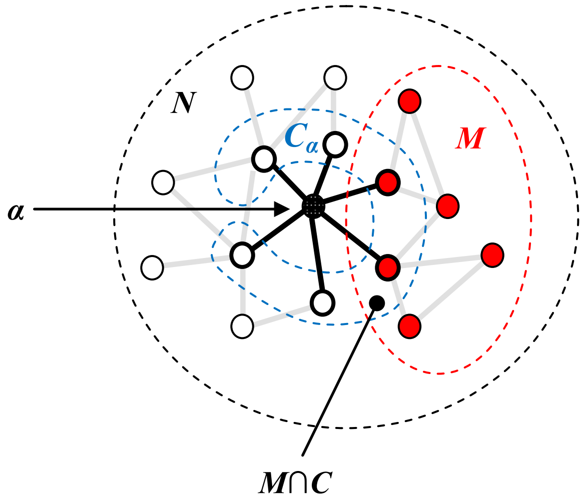
To examine the possibility of over-training of θ and θ′ values, we performed a leave -one-out cross-validation, in which each time one gene pair (l,k) was left out, the best θ and θ values were determined using the remaining gene pairs, and the left out gene pair was evaluated using these values. If $\rho(X^E{}_l,X^E{}_k)\geq\theta$ and $\rho(X^{E'}{}_l,X^{E'}{}_k)\geq\theta'$, the pair (l,k) was added to the cross-validation network G^x:

```
Set $G^x=\emptyset$
For all $\{(l,k)|l\in S',k\in S',(l,k)\in I\cup I'\}$
    If $(l,k)\in I$ then $I=I-\{(l,k)\}$
    If $(l,k)\in I'$ then $I'=I'-\{(l,k)\}$
    Find the values for $\theta$ and $\theta'$ using the new $I$ and $I'$
    If $\rho(X^E{}_l,X^E{}_k)\geq\theta$ and $\rho(X^{E'}{}_l,X^{E'}{}_k)\geq\theta'$ then $G^x= G^x+\{(l,k)\}$
    Restore $I$ and $I'$
Report $G^x$
```

The G^x was found to have a precision of 0.48 and S' coverage of 0.113 which are very close to the values for the conserved coexpression network that is reported in the paper, implying that the procedure used to find the best values for θ and θ′ did not over -train them.

8.4.3 Network-based prediction of gene function

We evaluated the association of each gene with each KEGG pathway using a hypergeometric-based method: Assume that N is the set of nodes in the network G, $C_\alpha\subset N$ is the set of nodes that are connected to the node α (excluding the node α itself), and $M\subset N$ is the set of nodes that have the particular function f^M according to KEGG, again excluding the node α itself:



The null hypothesis H_0 is that C_α is independent of M. To evaluate this hypothesis, we assume a hypergeometric distribution for $|M \cap C_\alpha|$:

$\Pr(X \geq x_{obs}|H_0) = \Sigma_x \text{ hypergeo}(x;|N|,|M|,|C_\alpha|)$,

where $x_{obs} = |M \cap C_\alpha| \leq x \leq \min(|M|,|C_\alpha|)$ and “hypergeo” is the hypergeometric distribution function. If H_0 is rejected, the node α is considered associated with M and, thus, with function f^M. Since node α itself is not included in the calculation of the probability value, there is no need to cross-validate this procedure, as it naturally resembles a leave-one-out cross-validated procedure.

We evaluated the performance of this procedure for each network and each pathway separately. The p-value cutoff for rejecting the null hypothesis was selected to be ≤ 0.05 and to result in a PPV≥ 0.80, meaning that at least 80% of predictions are correct.

8.4.4 Identification of potential regulatory motifs in UTRs

T. brucei genes were clustered based on the normalized values of the experiment set E_{KQJ}. We used different clustering approaches: Iclust [73] uses an information-based strategy to cluster the genes into a predefined number of clusters. By default, this number is $\sqrt{|S|}$, where S is the set of all *T. brucei* genes with available expression profiles. Alternatively, we used the standard k-means algorithm with either an initial set of 100 means or an initial set of 30 means. The algorithm converged to 82 and 19 clusters, respectively. Gene clusters along with either complete or truncated 3′UTR sequences were submitted to FIRE [168] with default parameters. The truncated sequences contained the first 1000bp from the 5′ end of each 3′ UTR. Prior to identification of potential regulatory elements, FIRE removes homologous sequences. In the paper, we only discuss the results of running FIRE on the set of 19 clusters and the truncated sequences; the complete set of results can be found at http://webpages.mcgill.ca/staff/Group2/rsalav/web/Suppl/20100109/index.htm.

8.5 Supplementary Tables

Supplementary Table 8-1. Gene-function associations based on the coexpression network CoExp$^{1}_{\mathbf{Tbr}}$ – Each gene was evaluated for its association with different KEGG pathways. The significant associations of genes that currently have no annotation in the KEGG pathway database are shown in this table. Old gene ids are brought in parentheses. Note that some of these genes, although not included in the KEGG pathway database, have functional annotations in TriTrypDB, most of which are congruent with the found association. An association with a pathway should be interpreted mainly as a functional linkage, meaning that the corresponding gene is either involved in that pathway, or has a function that is closely related to that pathway and, thus, requires synchronized expression.

	Oxidative phosphorylation	Purine metabolism	Ribosome	TriTrypDB annotation (v2.0)	Notes
Tb927.3.1410	*			Cytochrome oxidase subunit VII	
Tb09.160.1820	*			Cytochrome oxidase subunit V	
Tb927.5.1060	*			Mitochondrial processing peptidase, beta subunit	
Tb927.1.4100	*			Cytochrome oxidase subunit IV	
Tb927.10.9830 (Tb10.6k15.0480)	*			Hypothetical protein, conserved	*a*
Tb927.10.13360 (Tb10.389.0070)	*			Elongation factor Tu	
Tb927.5.2160	*			Hypothetical protein, conserved	
Tb927.4.720	*			Hypothetical protein, conserved	*b*
Tb927.10.11220 (Tb10.26.0790)	*			Procyclic form surface glycoprotein	
Tb927.10.15220 (Tb10.61.1290)	*			Hypothetical protein, conserved	
Tb11.01.1020	*			Hypothetical protein, conserved	
Tb927.10.15960 (Tb10.61.0320)	*			Hypothetical protein, conserved	
Tb927.4.4020		*		Amino acid transporter	*c*
Tb927.5.630		*		Acidic phosphatase	*d*
Tb927.10.7090 (Tb10.6k15.3640)		*		Alternative oxidase	*e*
Tb11.02.2430			***	60S ribosomal protein L17	
Tb11.01.5720			***	Ribosomal protein L18	
Tb927.8.1340			***	60S ribosomal protein l7a	

Tb927.6.5130			*	60S acidic ribosomal protein P2	
Tb11.02.4360			**	40S ribosomal protein S21	
Tb09.211.0340			*	60S ribosomal protein L10	

* $1\times10^{-4} <$ p-value ≤ 0.01

** $1\times10^{-7} <$ p-value $\leq 1\times10^{-4}$

*** $1\times10^{-14} <$ p-value $\leq 1\times10^{-7}$

a Tb10.6k15.0480 has been shown to be associated with ATP synthase complex [290].

b Tb927.4.720 has been reported as a mitochondrial protein [293].

c The metabolism of several amino acids, such as Ala, Asp, Glu, His, Gly, Ser and Thr is linked to purine metabolism pathway (KEGG pathway tbr00230). Also, the 3′ UTR of Tb927.4.4020 contains a copy of the motif [ACT]TAA[GT][AG][GT][CT][AC], which has been suggested as a potential purine metabolism-specific regulatory element [3].

d Acidic phosphatases are involved in purine nucleoside phosphate catabolism.

e Alternative oxidases are strongly regulated by purine nucleotides (see [296] for an example). In addition, the 3′ UTR of Tb10.6k15.3640 has an instance of the motif [ACGT][AGT][AC]TGCC[AG][GT] and several instances of the motif [ACT]TAA[GT][AG][GT][CT][AC], both of which are suggested as potential regulatory elements specific to purine metabolism [3].

Supplementary Table 8-2. Gene-function associations based on the coexpression network CoExp$^{2}_{\mathbf{Tbr}}$ – Pathways are defined according to KEGG. Refer to **Supplementary Table 8-1** for more details.

	Ribosome	Inositol phosphate metabolism	Phosphatidylinositol signaling	Glycolysis	TriTrypDB annotation (v2.0)	Notes
Tb927.6.2330	*****				RGG protein	*a*
Tb927.8.1110	*				40S ribosomal protein S9	
Tb927.8.1340	*****				60S ribosomal protein L7a	
Tb927.7.2370	****				40S ribosomal protein S15	
Tb11.01.5720	****				Ribosomal protein L18	
Tb927.6.5130	*				60S acidic ribosomal protein P2	
Tb11.02.4360	*****				40S ribosomal protein S21	
Tb11.02.2430	***				60S ribosomal protein L17	
Tb927.4.3660	***				Hypothetical protein, conserved	
Tb11.01.1465	***				Nascent polypeptide associated complex alpha	

					subunit	
Tb927.5.1110	*				60S ribosomal protein L2, 60S ribosomal protein L8	
Tb927.3.4360	**				40S ribosomal protein S15A	
Tb927.4.5030	*				Protein phosphatase 1	
Tb09.211.0340	****				60S ribosomal protein L10	
Tb11.01.7730	*				Hypothetical protein, conserved	*b*
Tb09.211.3300	*				Peroxin 19 (inferred from mutant phenotype)	*c*
Tb927.7.3530		*	*		Hypothetical protein, conserved	
Tb927.10.16170 (Tb10.61.0090)		*	*		Potassium voltage-gated channel	*d*
Tb11.01.3370			*		Glycosomal membrane protein	*e*
Tb927.7.4500				**	Hypothetical protein, conserved	
Tb927.4.4870				***	Amino acid transporter	

* $1\times10^{-4} <$ p-value≤ 0.01

** $1\times10^{-7} <$ p-value$\leq 1\times10^{-4}$

*** $1\times10^{-14} <$ p-value$\leq 1\times10^{-7}$

**** $1\times10^{-28} <$ p-value$\leq 1\times10^{-14}$

***** p-value$\leq 1\times10^{-28}$

a RGG domain has been shown previously to interact with ribosomal proteins [291].

b Tb11.01.7730 has been reported to be associated with transcription factor II H [297, 298].

c Tb09.211.3300 has several Pfam domains, such as Pex19 which is essential for peroxisome biogenesis, and LUC7 which is U1 snRNA-associated protein.

d The relationship between Inositol phosphate-mediated signaling and potassium channel activity has been reported in different studies [299-301].

e The 3′ UTR of Tb11.01.3370 contains the motifs [AT]CTTTT[GT]C[ACGT] and [ACG]AGAA[AC]A[AT][AGT]. Both of these motifs have been previously predicted as potential regulatory elements specific to inositol phosphate metabolism and phosphatidylinositol signaling genes [3].

Supplementary Table 8-3. Prediction of GO biological processes based on the coexpression network CoExp$^{1}_{Tbr}$ –Although some of these predictions correspond to genes that, according to TriTrypDB v2.0, are already annotated, they have not yet been assigned to any GO biological processes. Each prediction should be interpreted as a functional linkage, meaning that the corresponding gene either belongs to the predicted GO category, or has a function that is closely related to that category and, thus, requires synchronized expression.

	Glycolysis	rRNA processing	Antigenic variation	Protein folding	Cellular protein metabolic process	Regulation of cell cycle	TriTrypDB annotation (v2.0)
Tb927.1.4100	*						Cytochrome oxidase subunit IV
Tb927.10.1560		*					Hypothetical protein
Tb927.4.4790			*				Hypothetical protein
Tb927.3.2520			*				Expression site-associated gene (ESAG) protein
Tb927.3.2500			*				Hypothetical protein
Tb927.3.520			*				Expression site-associated gene (ESAG) protein
Tb927.3.3540				*	**	**	Nucleoporin
Tb927.7.5160					*	*	Deoxyuridine triphosphatase
Tb11.02.0080						*	Hypothetical protein
Tb927.4.1010						*	Hypothetical protein

* $1\times10^{-4} <$ p-value ≤ 0.01

** $1\times10^{-7} <$ p-value $\leq 1\times10^{-4}$

Supplementary Table 8-4. Prediction of GO molecular functions based on the coexpression network $\mathbf{CoExp^{1}_{Tbr}}$ – Refer to **Supplementary Table 8-3** for more details.

	Nucleic acid binding	ATP binding	Protein binding	Unfolded protein binding	TriTrypDB annotation (v2.0)
Tb09.211.1040	*				Hypothetical protein
Tb927.10.3910		*			Hypothetical protein
Tb927.7.5160		*			Deoxyuridine triphosphatase
Tb927.3.3540		*	*	*	Nucleoporin

* $1\times10^{-4} <$ p-value≤ 0.01

Supplementary Table 8-5. Prediction of GO cellular component based on the coexpression network $\mathbf{CoExp^{1}_{Tbr}}$ – Refer to **Supplementary Table 8-3** for more details.

	Cytoplasm	Chaperonin-containing T-complex	Plasma membrane	Cell surface	Integral to membrane	TriTrypDB annotation (v2.0)
Tb09.211.1220	*					Hypothetical protein
Tb927.3.1690	*					Hypothetical protein
Tb927.10.14790	*					Aminopeptidase
Tb927.7.5160		*				Deoxyuridine triphosphatase
Tb11.02.4750			*			Hypothetical protein
Tb09.211.3880			*			Hypothetical protein

Tb927.10.10000				**		Hypothetical protein
Tb927.4.1670				*		Hypothetical protein
Tb09.244.0640					*	Variant surface glycoprotein (VSG
Tb927.3.5690					*	Hypothetical protein
Tb11.01.7530					*	Hypothetical protein
Tb927.4.810					*	Expression site-associated gene (ESAG) protein
Tb927.10.6720					*	Hypothetical protein
Tb927.1.5160					*	Hypothetical protein
Tb927.10.5700					*	Hypothetical protein
Tb927.3.2520					*	Expression site-associated gene (ESAG) protein
Tb927.3.2500					*	Hypothetical protein
Tb927.5.310					*	Hypothetical protein
Tb927.5.1400					*	Hypothetical protein
Tb11.02.1564					*	Leucine-rich repeat protein (LRRP)
Tb927.3.1490					*	Leucine-rich repeat protein (LRRP)
Tb09.211.2060					*	Hypothetical protein
Tb927.8.5080					*	Hypothetical protein
Tb927.3.520					*	Expression site-associated gene (ESAG) protein

* 1×10^{-4} < p-value≤ 0.01

** 1×10^{-7} < p-value≤ 1×10^{-4}

Supplementary Table 8-6. Prediction of GO biological processes based on the coexpression network CoExp$^{2}_{Tbr}$ – Refer to **Supplementary Table 8-3** for more details.

	Translation	Ubiquitin-dependent protein catabolic process	Microtubule-based movement	ATP synthesis coupled proton transport	Antigenic variation	TriTrypDB annotation (v2.0)
Tb927.5.4120	**					Hypothetical protein
Tb09.160.2400	*					Hypothetical protein
Tb927.2.4700	*					Hypothetical protein

Tb927.10.12330	*					Hypothetical protein
Tb11.18.0014	*					Hypothetical protein
Tb927.4.3660	**					Hypothetical protein
Tb927.5.2520	*					Hypothetical protein
Tb927.6.2330	*****					RGG protein
Tb09.211.3300	*					Peroxin 19
Tb11.02.0445	*					Hypothetical protein
Tb927.4.1340	**					Cleavage and polyadenylation specificity factor subunit
Tb11.01.8690	*					Hypothetical protein
Tb927.4.5030	**					Protein phosphatase 1
Tb927.6.2210	*					Hypothetical protein
Tb09.160.3160	*					Hypothetical protein
Tb927.7.640	*					Hypothetical protein
Tb927.8.5990	*					Hypothetical protein
Tb11.01.7730	*					Hypothetical protein
Tb927.7.4120	*					Hypothetical protein
Tb927.3.1370	***					40S ribosomal protein S25
Tb927.10.12310		*				Helicase-like protein
Tb09.211.1270			*			Hypothetical protein
Tb11.02.3880			*			Hypothetical protein
Tb11.02.4120				**		Hypothetical protein
Tb927.3.2180				**		Hypothetical protein
Tb09.142.0310					*	Expression site-associated gene (ESAG) protein
Tb09.211.4820					**	Hypothetical protein
Tb927.10.8980					***	Hypothetical protein
Tb09.160.1440					**	Hypothetical protein
Tb11.02.1470					***	Hypothetical protein
Tb927.4.990					**	Hypothetical protein
Tb11.38.0003					***	Variant surface glycoprotein (VSG)
Tb927.3.5830					***	Expression site-associated gene (ESAG) protein
Tb927.1.5030					**	Leucine-rich repeat protein (LRRP)
Tb09.244.1950					***	Hypothetical protein
Tb927.5.750					**	Hypothetical protein
Tb927.3.2520					**	Expression site-associated gene (ESAG) protein
Tb927.10.5700					*	Hypothetical protein
Tb927.5.390					**	75 kDa invariant surface glycoprotein
Tb11.01.7860					*	Hypothetical protein
Tb927.3.1870					**	Hypothetical protein
Tb927.6.540					***	Gene related to expression site-associated gene 2 (GRESAG2)

*	$1\times10^{-4} <$ p-value ≤ 0.01
**	$1\times10^{-7} <$ p-value $\leq 1\times10^{-4}$
***	$1\times10^{-14} <$ p-value $\leq 1\times10^{-7}$
****	$1\times10^{-28} <$ p-value $\leq 1\times10^{-14}$
*****	p-value $\leq 1\times10^{-28}$

Supplementary Table 8-7. Prediction of GO molecular functions based on the coexpression network CoExp$^{2}_{\mathrm{Tbr}}$ – Refer to **Supplementary Table 8-3** for more details.

	Structural constituent of ribosome	Microtubule motor activity	Catalytic activity	TriTrypDB annotation (v2.0)
Tb927.5.4120	**			Hypothetical protein
Tb09.160.2400	*			Hypothetical protein
Tb927.2.4700	*			Hypothetical protein
Tb927.4.3660	**			Hypothetical protein
Tb927.10.3970	*			Hypothetical protein
Tb11.02.0445	*			Hypothetical protein
Tb09.160.3160	*			Hypothetical protein
Tb11.01.7730	*			Hypothetical protein
Tb927.7.4120	*			Hypothetical protein
Tb927.3.1370	***			40S ribosomal protein S25
Tb09.211.1270		*		Hypothetical protein
Tb11.02.3880		*		Hypothetical protein
Tb927.3.3690		*		Flagellar radial spoke protein-like
Tb11.46.0005			**	Hypothetical protein

*	$1\times10^{-4} <$ p-value ≤ 0.01
**	$1\times10^{-7} <$ p-value $\leq 1\times10^{-4}$
***	$1\times10^{-14} <$ p-value $\leq 1\times10^{-7}$

Supplementary Table 8-8. Prediction of GO cellular component based on the coexpression network CoExp$^{2}_{\mathbf{Tbr}}$ – Refer to **Supplementary Table 8-3** for more details.

	Ribosome	Intracellular	Cytoplasm	Chaperonin-containing T-complex	Integral to membrane	Dynein complex	TriTrypDB annotation (v2.0)
Tb11.02.0445	*						Hypothetical protein
Tb927.5.4120	**	**					Hypothetical protein
Tb09.160.2400	*	*					Hypothetical protein
Tb927.2.4700	*	*					Hypothetical protein
Tb927.3.3570	*	*					Hypothetical protein
Tb927.4.3660	**	**					Hypothetical protein
Tb927.6.2330	*****	*****					RGG protein
Tb927.10.3970	*	*					Hypothetical protein
Tb927.6.1130	*	*					Hypothetical protein
Tb09.211.4690	*	*					Hypothetical protein
Tb09.160.3160	*	*					Hypothetical protein
Tb927.8.6250	*	*					Hypothetical protein
Tb11.01.7730	*	*					Hypothetical protein
Tb927.10.160	*	*					Hypothetical protein
Tb11.01.1570	*	*					NUDIX hydrolase
Tb927.7.4120	*	*					Hypothetical protein
Tb927.3.1370	***	***					40S ribosomal protein S25
Tb11.55.0016		*					Hypothetical protein
Tb927.7.6280		*					Hypothetical protein
Tb927.7.3530		*					Hypothetical protein
Tb927.7.3580		*					Protein kinase
Tb09.244.2390		*					Hypothetical protein
Tb11.02.0450		*					Hypothetical protein
Tb927.10.12100		*					RNA-binding protein
Tb927.3.3880		*					Hypothetical protein
Tb927.4.3830		*					Hypothetical protein

Tb927.10.9050		*					Hypothetical protein
Tb09.211.4180		*					Hypothetical protein
Tb927.5.4010		*					Hypothetical protein
Tb11.01.3530		*					Hypothetical protein
Tb927.4.2570		*					Hypothetical protein
Tb927.2.5440		*					Hypothetical protein
Tb09.244.2710		*					Hypothetical protein
Tb09.160.2900		*					PRP3
Tb11.01.6000		*					Hypothetical protein
Tb11.02.4790		*					ATG16/SAP18/CVT11/APG16
Tb09.211.0500		*					Hypothetical protein
Tb927.10.16140		*					Adenylate/guanylate cyclase
Tb927.3.2470		*					Pumilio RNA binding protein
Tb927.7.5150		*					Hypothetical protein
Tb927.10.9180		*					Hypothetical protein
Tb927.1.1400		*					Hypothetical protein
Tb11.01.1910		*					Hypothetical protein
Tb927.7.1450		*					Hypothetical protein
Tb927.6.1180		*					Hypothetical protein
Tb927.4.1980		*					Hypothetical protein
Tb927.1.2200		*					Hypothetical protein
Tb927.8.3090		*					Hypothetical protein
Tb927.5.1770		*					Hypothetical protein
Tb927.8.2000		*					Cyclophilin type peptidyl-prolyl cis-trans isomerase
Tb11.01.6835		*					Hypothetical protein
Tb927.2.2450		*					Ribosomal RNA methyltransferase
Tb927.7.640			**				Hypothetical protein
Tb09.160.2090				*			Hypothetical protein
Tb927.5.2570				*			Translation initiation factor
Tb09.211.1360				*			Hypothetical protein
Tb09.211.0690				**			Hypothetical protein
Tb927.10.14790				*			Aminopeptidase
Tb09.244.2190					*		Hypothetical protein
Tb09.v1.0820					*		Hypothetical protein
Tb927.5.3100					*		Hypothetical protein
Tb927.10.9510					***		Hypothetical protein
Tb09.211.4820					*		Hypothetical protein
Tb09.160.5350					*		Variant surface glycoprotein (VSG)-related
Tb09.142.0320					*		Hypothetical protein
Tb11.02.1564					*		Leucine-rich repeat protein (LRRP)
Tb927.10.1770					**		Hypothetical protein

Tb11.02.3710					*		Hypothetical protein
Tb11.02.1565					*		Hypothetical protein
Tb927.10.6740					*		Hypothetical protein
Tb927.1.5160					***		Hypothetical protein
Tb927.1.5180					*		Hypothetical protein
Tb927.3.580					**		Leucine-rich repeat protein (LRRP)
Tb11.02.1640					*		Kinetoplastid-specific dual specificity phosphatase
Tb11.01.6210					*		Procyclin-associated gene 2-like protein
Tb927.3.2590					*		Hypothetical protein
Tb927.3.570					*		Expression site-associated gene (ESAG) protein
Tb09.244.0640					***		Variant surface glycoprotein (VSG
Tb927.10.8980					**		Hypothetical protein
Tb11.01.6140					*		Hypothetical protein
Tb927.1.1850					*		Hypothetical protein
Tb09.160.1440					*		Hypothetical protein
Tb11.02.1470					**		Hypothetical protein
Tb927.10.1230					**		Hypothetical protein
Tb927.3.3980					*		Hypothetical protein
Tb927.10.14770					*		Protein kinase
Tb927.8.7540					*		Hypothetical protein
Tb927.4.4790					*		Hypothetical protein
Tb927.10.3360					*		Hypothetical protein
Tb11.01.6220					*		Procyclin-associated gene 4 (PAG4) protein
Tb927.10.1040					*		Serine carboxypeptidase III precursor
Tb927.4.990					*		Hypothetical protein
Tb927.7.380					*		Hypothetical protein
Tb11.38.0003					**		Variant surface glycoprotein (VSG)
Tb927.6.1730					*		Hypothetical protein
Tb927.10.8860					*		Hypothetical protein
Tb09.160.4760					*		Hypothetical protein
Tb927.3.5830					***		Expression site-associated gene (ESAG) protein
Tb927.1.5030					*		Leucine-rich repeat protein (LRRP)
Tb927.10.9450					**		Hypothetical protein
Tb927.5.1440					*		Hypothetical protein
Tb927.10.530					*		Hypothetical protein
Tb927.7.190					*		Thimet oligopeptidase A
Tb927.4.1110					*		Hypothetical protein
Tb927.5.3600					*		ATP-dependent DEAD/H RNA helicase
Tb09.244.1950					***		Hypothetical protein
Tb09.160.0360					*		Hypothetical protein
Tb11.01.7380					*		Hypothetical protein

Tb927.3.1490					*		Leucine-rich repeat protein (LRRP)
Tb11.01.7530					*		Hypothetical protein
Tb927.8.980					*		Phosphoacetylglucosamine mutase
Tb927.1.5060					*		Variant surface glycoprotein (VSG)-related
Tb927.5.750					*		Hypothetical protein
Tb927.4.810					*		Expression site-associated gene (ESAG) protein
Tb927.10.1780					*		Hypothetical protein
Tb927.3.2520					**		Expression site-associated gene (ESAG) protein
Tb927.3.5720					*		Hypothetical protein
Tb927.5.4600					**		Expression site-associated gene (ESAG) protein
Tb927.5.1400					***		Hypothetical protein
Tb927.3.560					**		Expression site-associated gene (ESAG) protein
Tb927.10.15440					*		Hypothetical protein
Tb927.3.980					**		Hypothetical protein
Tb927.10.5710					*		Hypothetical protein
Tb927.10.5700					*		Hypothetical protein
Tb927.5.1390					*		64 kDa invariant surface glycoprotein
Tb927.7.6860					*		Expression site-associated gene (ESAG) protein
Tb927.3.2500					*		Hypothetical protein
Tb927.8.4360					*		Hypothetical protein
Tb927.3.5680					*		Variant surface glycoprotein (VSG)-related
Tb11.01.7860					*		Hypothetical protein
Tb927.3.1870					**		Hypothetical protein
Tb927.2.3340					**		Hypothetical protein
Tb927.8.6720					*		Hypothetical protein
Tb927.8.5080					*		Hypothetical protein
Tb927.8.7330					**		Hypothetical protein
Tb927.3.520					**		Expression site-associated gene (ESAG) protein
Tb927.3.5690					***		Hypothetical protein
Tb927.8.7310					*		Hypothetical protein
Tb927.6.540					**		Gene related to expression site-associated gene 2 (GRESAG2)
Tb927.1.2600						*	Pumilio/PUF RNA binding protein 9
Tb09.160.0720						*	Hypothetical protein
Tb11.01.5260						*	Radial spoke protein RSP11
Tb927.1.2760						*	Hypothetical protein
Tb09.244.2050						*	Hypothetical protein
Tb09.244.1650						*	Hypothetical protein
Tb11.01.6390						*	Hypothetical protein
Tb927.3.4510						*	Hypothetical protein
Tb927.10.8780						*	Hypothetical protein

Tb927.1.4310						*	Hypothetical protein
Tb11.02.3880						**	Hypothetical protein
Tb927.3.3300						*	Hypothetical protein
Tb11.02.4640						*	Tubulin-tyrsoine ligase-like protein
Tb927.3.3110						*	Hypothetical protein

* $1\times10^{-4} <$ p-value ≤ 0.01

** $1\times10^{-7} <$ p-value $\leq 1\times10^{-4}$

*** $1\times10^{-14} <$ p-value $\leq 1\times10^{-7}$

**** $1\times10^{-28} <$ p-value $\leq 1\times10^{-14}$

***** p-value $\leq 1\times10^{-28}$

Supplementary Table 8-9. Gene-function associations based on the conserved coexpression network CoExp$_{\mathbf{Tbr\times Lif}}$ – Pathways are defined according to KEGG. Refer to **Supplementary Table 8-1** for more details.

	Proteasome	Oxidative phosphorylation	Ribosome	TriTrypDB annotation (v2.0)	Notes
Tb927.10.6800 (Tb10.6k15.3970)	*			Developmentally regulated GTP-binding protein	
Tb927.5.1520	*			Heat shock protein HslVU, ATPase subunit HslU	
Tb927.7.4870	**			Hypothetical protein, conserved	
Tb09.160.0740		*		Hypothetical protein, conserved	
Tb927.10.9830 (Tb10.6k15.0480)		*		Hypothetical protein, conserved	*a*
Tb927.10.8320 (Tb10.6k15.2180)		*		Cytochrome oxidase subunit IX	
Tb11.02.4120		**		Hypothetical protein, conserved	*a*
Tb09.160.1820		*		Cytochrome oxidase subunit V	
Tb927.10.5050 (Tb10.70.2155)		*		Hypothetical protein, conserved	*b*
Tb927.10.4240 (Tb10.70.3010)		**		Hypothetical protein, conserved	*c,d*
Tb927.3.1410		*		Cytochrome oxidase subunit VII	
Tb927.3.700		*		Hypothetical protein, conserved	*d*
Tb927.4.3450		*		Hypothetical protein, conserved	*b,d*
Tb927.8.5120		*		Cytochrome c	
Tb927.10.520 (Tb10.70.7760)		*		Hypothetical protein, conserved	*a*

Tb927.5.1060		**		Mitochondrial processing peptidase, beta subunit	
Tb927.5.2930		**		Hypothetical protein, conserved	*a*
Tb927.4.1360		*		Hypothetical protein, conserved	*e*
Tb11.47.0022		*		Hypothetical protein, conserved	*a*
Tb927.6.590		*		Hypothetical protein, conserved	*b*
Tb927.7.3500		*		Glutathione-S-transferase/glutaredoxin	
Tb927.7.840		*		Hypothetical protein, conserved	*a*
Tb927.5.3090		*		Hypothetical protein, conserved	*a*
Tb927.10.15960 (Tb10.61.0320)		*		Hypothetical protein, conserved	
Tb927.4.720		*		Hypothetical protein, conserved	*b*
Tb927.3.920		*		Hypothetical protein, conserved	
Tb927.5.590			*	Protein phosphatase 1, regulatory subunit	*f*
Tb11.01.5720			*	Ribosomal protein L18	
Tb11.55.0002			*	Protein phosphatase 2C	
Tb927.4.3660			*	Hypothetical protein, conserved	

* $1\times10^{-4} <$ p-value ≤ 0.01

** $1\times10^{-7} <$ p-value $\leq 1\times10^{-4}$

a These proteins have been shown to be associated with ATP synthase complex [290].

b These proteins have been reported to be mitochondrial [293].

c These proteins have been reported to be associated with mitochondrial membrane [294].

d The mRNA of these proteins have the UAG(G)UA(G/U) element, which is also found in the transcripts of many cytochrome c oxidase subunits [61].

e Tb927.4.1360 is suggested as a PF-specific glycosomal protein [302].

f Protein phosphatase 1 is known to regulate the activity of ribosomal protein S6 [303].

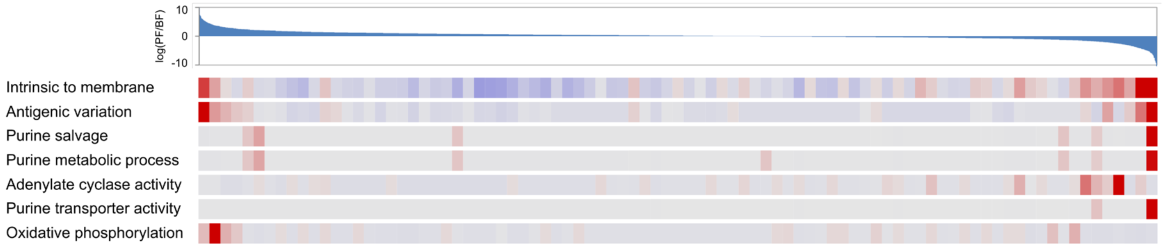
Supplementary Table 8-10. Prediction of GO terms based on the conserved coexpression network CoExp$_{\mathrm{Tbr\times Lif}}$ – Refer to **Supplementary Table 8-3** for more details.

	Biological Process		Molecular Function	Cellular Compartment		
	Cellular protein metabolic process	Regulation of cell cycle	Catalytic activity	Chaperonin-containing T-complex	Ribosome	TriTrypDB annotation (v2.0)
Tb11.46.0009	**	**		**		Hypothetical protein
Tb927.3.1940			*			Hypothetical protein
Tb927.5.3090			*			Hypothetical protein
Tb927.1.1390			*			Hypothetical protein
Tb927.5.590					*	Protein phosphatase 1

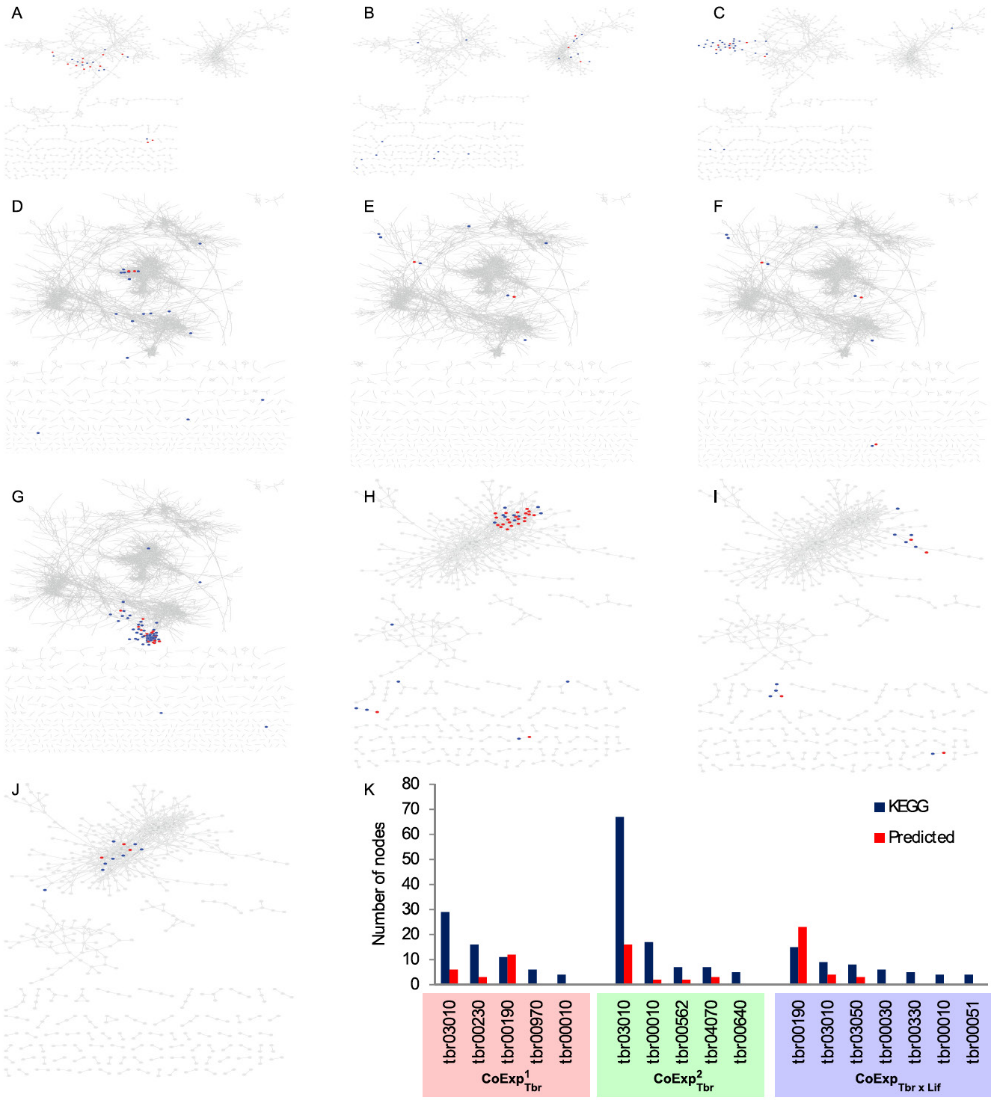
* 1×10^{-4} < p-value ≤ 0.01

** 1×10^{-7} < p-value ≤ 1×10^{-4}

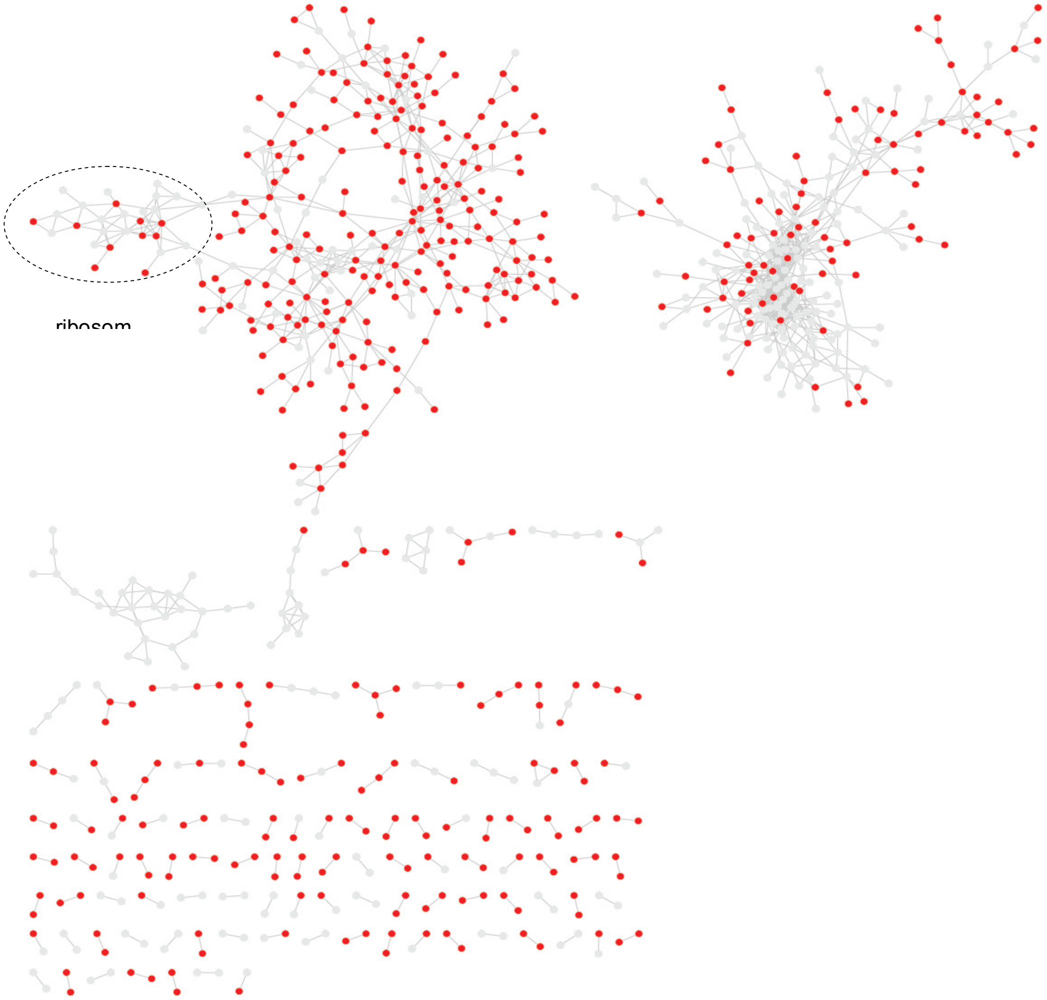
8.6 Supplementary Figures



Supplementary Figure 8-1. Functions that are over-expressed in PF or BF *T. brucei* –Each row indicates a particular category according to either GO or KEGG, and each column represents a set of genes, whose relative expression in PF and BF cells is indicated in the graph above. Red and blue colors indicate over-representation and under-representation, respectively. Over- and under-representation were calculated based on hypergeometric distribution assumption for the overlap of each functional category with each expression bin. Some categories such as proteins that are intrinsic to membrane or proteins that are involved in antigenic variation are enriched in both PF-specific (left) and BF-specific (right) genes, while metabolism and transport of purines and adenylate cylase activity are mainly over-represented among BF-specific genes, and oxidative phosphorylation is expressed only in PF.



Supplementary Figure 8-2. Distribution of different KEGG pathways in *T. brucei* coexpression and conserved coexpression networks –(A) oxidative phosphorylation (tbr00190) proteins in CoExp$^1_{\text{Tbr}}$; **(B)** purine metabolism (tbr00230) proteins in CoExp$^1_{\text{Tbr}}$; **(C)** ribosome (tbr03010) proteins in CoExp$^1_{\text{Tbr}}$; **(D)** glycolysis/gluconeogenesis (tbr00010) proteins in CoExp$^2_{\text{Tbr}}$; **(E)** inositol phosphate metabolism (tbr00562) proteins in CoExp$^2_{\text{Tbr}}$; **(F)** phosphatidylinositol signaling system (tbr04070) proteins in CoExp$^2_{\text{Tbr}}$; **(G)** ribosome (tbr03010) proteins in CoExp$^2_{\text{Tbr}}$; **(H)** oxidative phosphorylation (tbr00190) proteins in CoExp$_{\text{Tbr}\times\text{Lif}}$; **(I)** ribosome (tbr03010) proteins in CoExp$_{\text{Tbr}\times\text{Lif}}$; **(J)** proteasome (tbr03050) proteins in CoExp$_{\text{Tbr}\times\text{Lif}}$. **(K)** This graph shows the number of proteins that are annotated in KEGG, or whose annotation is predicted, in each network. Only KEGG pathways that could be predicted with at least 80% precision are shown. Note that for some of these pathways no new gene could be predicted.



Supplementary Figure 8-3. **Distribution of conserved proteins in CoExp1Tbr –**The orthologs of many *T. brucei* proteins with known KEGG pathways cannot be unambiguously identified in *L. infantum*; this is particularly the case for ribosomal proteins whose sequence similarity to each other prevents unambiguous identification of orthologous partners. Proteins whose *L. infantum* orthologs can be unambiguously identified are shown by red nodes.

9 Global identification of conserved post-transcriptional regulatory programs in trypanosomatids

In the previous chapter, we showed that based on expression patterns across different life stages, biological processes and pathways can be predicted in *T. brucei*. In other words, while the expression patterns of genes that are in the same pathway are similar to each other, expression patterns of different pathways are different. This suggests the presence of a complex post-transcriptional regulatory network in *T. brucei*, which is capable of producing the observed variability in the expression patterns of different genes. In this chapter, we present a comprehensive analysis of *T. brucei* 3' UTRs for identification of conserved *cis*-regulatory elements. We have developed a new statistical framework that uses the notion of network-level conservation for identification of conserved regulatory programs across multiple species. This framework provides us with an alignment-free method for identification of conserved *cis*-regulatory elements, and in particular is able to integrate structure and sequence information in order to identify potentially structural RNA motifs. We present a thorough analysis of available microarray and RNA-seq data in order to validate these conserved motifs, and propose potential *trans*-acting proteins that can bind to and modulate the mRNAs that contain these motifs. This study provides a universal framework for identification of conserved regulatory programs across multiple species, and introduces the first global map of post-transcriptional regulation in trypanosomatids.

9.1 Background

Identification of the mechanisms that regulate cellular processes is crucial to understanding the cell development and behavior. Regulatory networks are considerably more complex in eukaryotes than in prokaryotes, consisting of myriads of regulatory interactions among proteins, RNA molecules, and genomic DNA. Markedly, post-transcriptional events play a major role in the regulation of eukaryotic genes. While transcriptional regulation of gene expression has been the subject of many studies over the years [74, 304], the widespread role of post-transcriptional events in gene regulation has rather recently come to light [305-308]. Post-transcriptional regulation primarily involves the interaction of a *cis*-regulatory RNA element and a *trans*-acting element, which is usually either a microRNA (miRNA) or an RNA-binding protein.

Unlike the *cis*-acting elements that are involved in transcriptional regulation, most *cis*-acting post-transcriptional regulatory elements are located in the mRNA untranslated region (UTR), many of which form distinct secondary structures that are specifically recognized by their *trans*-acting binding partners [309-311]. Several studies have used computational methods to address the problem of identifying the structural RNA motifs that are involved in post-transcriptional regulation. These methods identify structural RNA regulatory elements based on commonality in a set of related sequences [312], the ability to explain expression data [313], or conservation across species [314-317]. Many conservation-based methods require an aligned set of RNA sequences that contain the putative regulatory element(s) [318]. These methods have limited applicability when the sequences are highly divereged and cannot be aligned reliably. In contrast, methods that identify structural RNA motifs from unaligned sequences assume that, in homologous sequences, similarity is limited to functional parts, such as regulatory elements [314]. Therefore, these methods can identify structural motifs even within sets of sequences that cannot be aligned. Nonetheless, these methods consider only the conservation of *cis*-regulatory elements, discarding the useful information that the conservation of the regulatory "network" provides.

“Network-level” conservation of regulatory programs implies that if A and Á are two orthologous *trans*-acting regulatory elements in two different organisms, the target genes of A (i.e. the genes that are bound and regulated by A) are mostly the orthologs of the target genes of Á. Network-level conservation, combined with the observation that the binding preferences of *trans*-acting regulatory elements are conserved across species, has been successfully used to identify conserved linear *cis*-regulatory motifs in pairs of organisms [319, 320].

Here, we introduce a general framework for identification of linear and structural motifs based on network-level conservation across multiple species. We employ this framework in order to identify conserved post-transcriptional regulatory programs in trypanosomatids, a group of organisms in which gene regulation is mainly at the post-transcriptional level. The responsiveness of this regulatory network to developmental events as well as external and internal stimuli indicates that trypanosomatid genes are modulated by a complex set of *cis*- and *trans*-regulatory elements at the post-transcriptional level within and across different life stages.

9.2 Results and Discussion

9.2.1 Identification of linear and structural regulatory motifs that are conserved at the network level

The statistical framework that we have developed allows us to measure the network-level conservation of potential regulatory motifs across multiple species (Supplementary Figure 9-1). Using this approach, we searched in a set of about ~4.7×10^6 linear and structural motifs (see the Methods section) in order to identify 3’ UTR regulatory elements that are conserved at the network level across different species of the genus of *Trypanosoma*, including *T. brucei*, *T. cruzi*, *T. congolense*, and *T. vivax*. These organisms are known to regulate their genes post-transcriptionally, using *cis*-regulatory factors that are mainly located in the mRNA 3’ UTRs [3, 5, 321, 322].

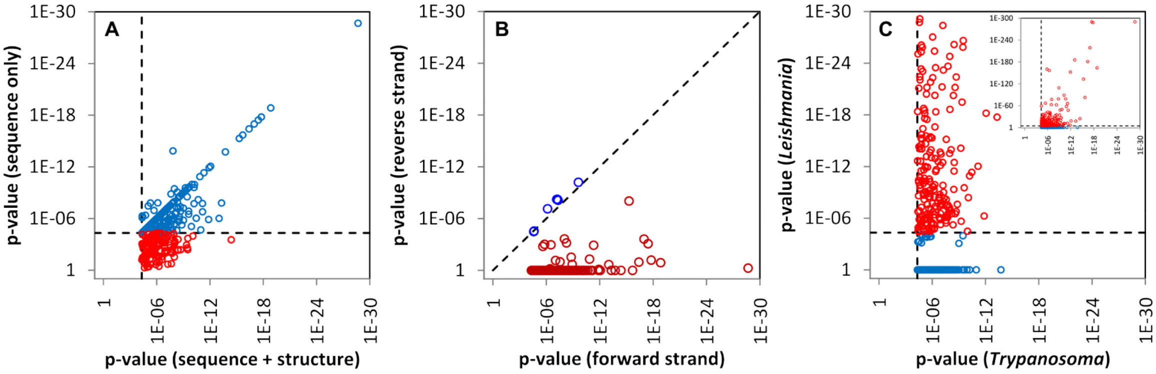


Figure 9-1. Network-level conservation identifies RNA motifs with structural and sequence information – We identified 388 linear and structural non-redundant motifs with significant network-level conservation across the genus of *Trypanosoma*. **(A)** When the structural information is discarded, 222 motifs remain significant for their network-level conservation (blue), while the conservation scores of 166 motifs drop below the significane threshold (red). The latter motifs are deemed to have indispensible structural information. **(B)** Almost all motifs are exclusively conserved in the forward strand, and not in the reverse strand. The few motifs whose p-values are similar in the forward and reverse strand (blue) contain palindromic sequences. **(C)** Of the 388 motifs that are conserved in the genus of *Trypanosoma*, 237 are also conserved in the genus of *Leishmania* (red). For visualization purposes, the main graph displayes motifs whose p-values in *Leishmania* are greater than 10^{-30}; the inset graph represents all motifs. Note the reverse order of p-values in the graphs, with smaller (significant) p-values toward the right/top of each chart.

After removal of redundant motifs, our search resulted in 388 putative regulatory elements that were highly conserved in the *Trypanosoma* genus with an estimated false discovery rate (FDR) of ~0.01 (Supplementary Table 9-1). The structural information of 166 of these motifs is indispensable, meaning that if the structural information was discarded and only the sequence information was retained, these motifs would not be identified as conserved anymore (Figure 9-1A). Furthermore, as expected from RNA *cis*-regulatory elements, almost all of the identified motifs are only conserved on the forward strand of the DNA, and show little or no network-level conservation when the reverse strand is considered (Figure 9-1B). We also found that the motifs that we identified are conserved not only within the genus of *Trypanosoma*, but also within another branch of trypanosomatids, namely the genus of *Leishmania*, which includes parasitic species such as *L. major*, *L. infantum*, *L. braziliensis*, and *L. mexicana* (Figure 9-1C). This suggests that these motifs have conserved their function beyond the *Trypanosoma* genus, and that the corrsponding regulatory network is conserved at least across the order of *Trypanosomatida*. Also, functional interactions among several motifs of this network suggest modularity (Supplementary Figure 9-2), which has been proposed to confer robustness and rapid responsiveness [323].

Several of the motifs that we identified match already known *cis*-regulatory elements that are involved in post-transcriptional regulation of different transcripts in many organisms, including trypanosomatids. For example, the most conserved motif, CAUAGAN, matches the known binding site of the trypanosomatid cycling sequence binding proteins (CSBPs), which determine the stability of S phase-specific transcripts [324]. Also, we were able to identify a structural motif that matches the well-studied histon 3' UTR stem-loop, consisting of a six-base pair stem and a four-nucleotide loop with the consensus sequence NGCUNUUNNNRNRGYN (the stem region is underlined). This motif is involved in transport and regulation of histon transcripts [325]. As another example, we identified a highly conserved linear motif with the consensus sequence AUGUAN. This motif contains the core binding sequence of the PUF family of RNA-binding proteins [326]. A similar motif has been previously reported to be over-represented among several groups of co-regulated transcripts in trypanosomatids [327].

These few examples suggest that our statistical framework can successfully identify conserved regulatory elements that are involved in post-transcriptional gene regulation. In order to systematically validate the discovered motifs, we further analyzed them by examining their profile across several microarray and RNA-seq experiments, as described in the next section.

9.2.2 The conserved post-transcriptional regulatory network of trypanosomatids is correlated with mRNA abundance

T. brucei is one of the major disease-causing trypanosomatids, responsible for the deadly human African trypanosomiasis, also known as sleeping sickness. The life cycle of *T. brucei* mainly consists of the insect stage, dominated by procyclic form (PF) parasites, and the mammalian stage, dominated by bloodstream form (BF) parasites. The transcriptome of *T. brucei* has been profiled through the life cycle as well as in different genetic backgrounds. By Analyzing available microarray and RNA-seq data of *T. brucei*, we found that several of the conserved RNA motifs that we have identified are potentially involved in regulation of genes through the life cycle of this organism. Specifically, 22 motifs show significant up-regulation or down-regulation in at least one experiment (Figure 9-2). For example, the AU-rich element AUUUAUU, designated in this article as Ptrm1970 (post-transcriptional regulatory motif 1970), is highly enriched among transcripts that are up-regulated in the stationary-phase *in vitro*-cultured PF *T. brucei* as well as in transcripts that are down-regulated in the stumpy and slender BF *T. brucei*. On the other hand, transcripts that contain the highly conserved linear motif UYGCNGA (Ptrm23) are down-regulated in the stationary-phase PF and up-regulated in the slender BF parasites. Such motifs may be involved in the progression of life cycle and differentiation of the parasite as well as regulating the biological processes that are specifically required in each life stage. However, this may not necessarily be the case for all these motifs. For example, up-regulation of the CAUAGAN motif (Ptrm1) in the log-phase PF cells may simply reflect the high cell growth rate and thus the high abundance of S phase parasites at the log phase, leading to the domination of the extracted mRNA pool by S phase-specific transcripts that contain this motif. Therefore, although analysis of the expression data supports these motifs as genuine *cis*-regulatory elements, interpreting the biological role of these regulatory elements based on the expression data is not a trivial task and needs orthogonal information.

In addition to analyzing individual microarray/RNA-seq experiments, we analyzed the expression profiles of *T. brucei* transcripts across different experiments in order to identify conserved motifs that occur in sets of co-regualted mRNAs. We identified 24 motifs that had significant local enrichment in the gene expression hyperspace of *T. brucei* (Figure 3), eight of which could only be identified by this cross-experiment analysis and not by the enrichment analysis of single experiments.

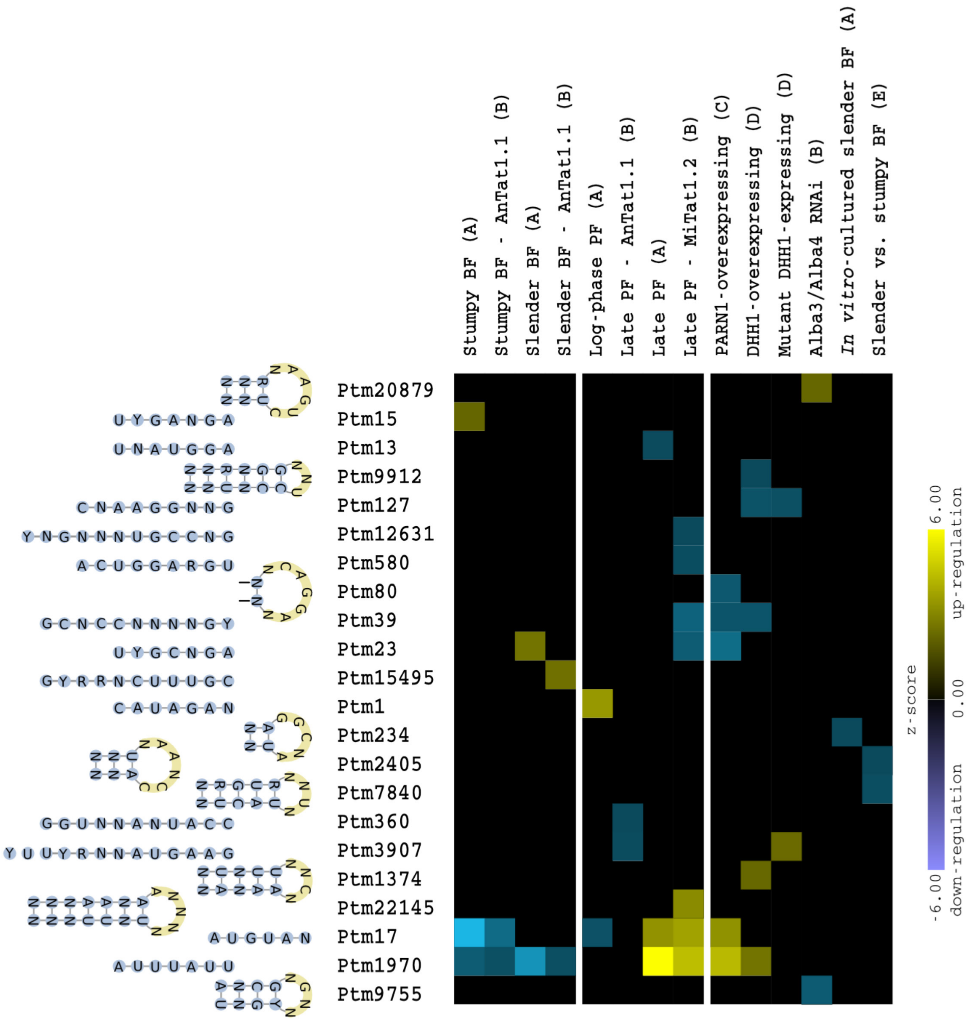


Figure 9-2. Conserved RNA motifs correlate with available microarray and RNA-seq data of *T. brucei* – Based on Mann-Whitney U test, 23 RNA motifs are significantly up-regulated (yellow) or down-regulated (blue) in at least one available expression dataset of *T. brucei*. The motif name along with the structure/sequence is shown on the left, with each column representing one expression dataset. A motif is deemed structural if combination of sequence and structure results in a better conservation *p*-value compared to sequence alone; otherwise the motif is deemed linear. A yellow square indicates significant up-regulation of the corresponding motif in the respective expression dataset, while a blue square indicates significant down-regulation. The letters in brackets correspond to the reference publications that describe each experiment: A: [35], B: [38], C: [37], D: [33], E: [34].

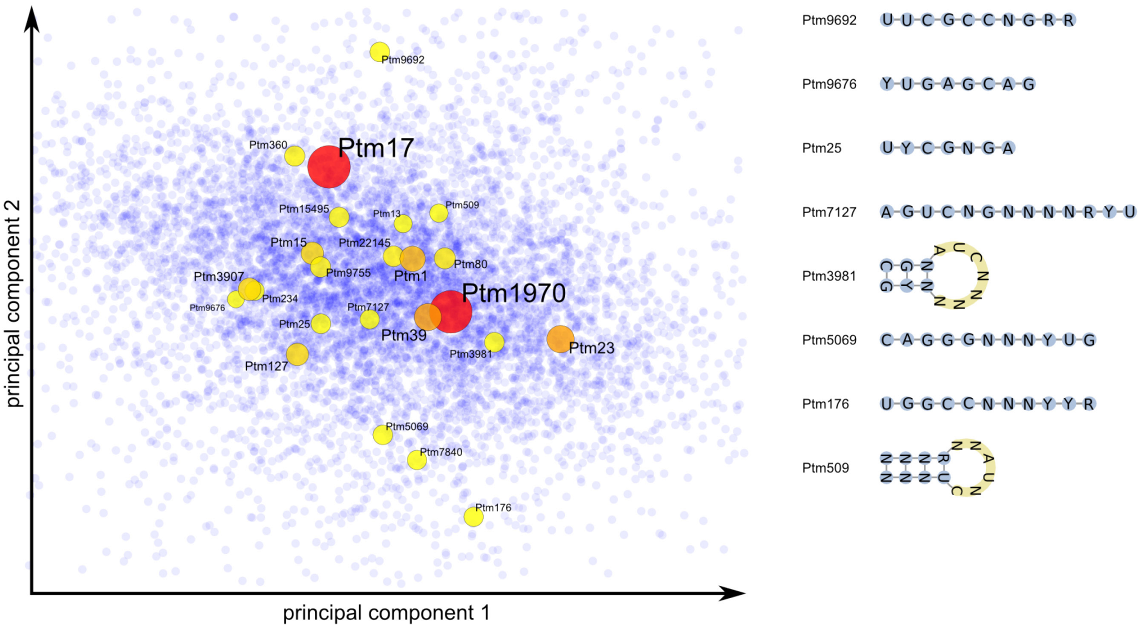


Figure 9-3. Transcripts that contain similar conserved RNA motifs are co-regulated across different conditions and life stages – *T. brucei* genes (blue dots) are mapped on the first two principal components of 22 previously published expressoin datasets [33-39]. Local enrichment of motifs were examined in different regions of the expression hyperspace (see the Methods section). Regions with significant local enrichment for motifs are highlighted in this figure by the circles. Larger/red circles represent higher enrichment z-scores. Seven novel motifs were validated using this analysis in comparison to single-array analysis, the sequence/structure of which is shown on the right.

9.2.3 A high-confidence gene regulatory network (GRN) suggests major regulatory role for two conserved cis-acting elements in Trypanosoma brucei

We constructed a high-confidence GRN of *T. brucei* based on the assumption that the transcripts that carry true instances of the same *cis*-regulatory motif must be co-regulated across different conditions. Therefore, the true targets of a particular *trans*-regulatory element are transcripts that contain the corresponding *cis*-regulatory motif and are co-expressed with other carriers of that motif. This high-confidence GRN contains 1012 interactions between 917 genes and 12 unknown *trans*-regulatory elements (Figure 9-4). The *trans*-regulatory elements that bind to Ptrm17 (AUGUAN) and Ptrm1970

(AUUUAUU) dominate the regulatory interactions of this GRN, targeting 508 and 328 genes, respectively. In the next sections, we examine the specific functions of these motifs.

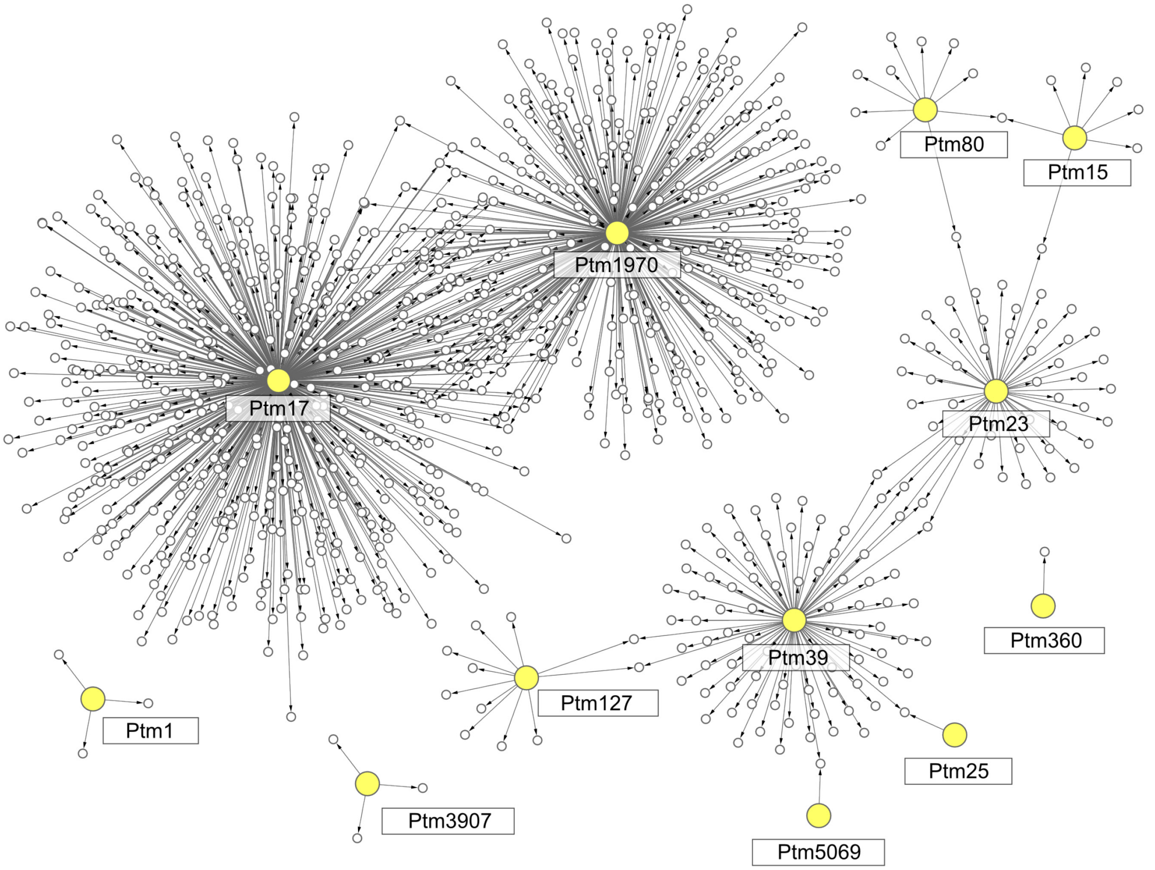


Figure 9-4. The high-confidence GRN of *T. brucei* – For each motif, a hypothetical unknown *trans*-regulatory factor is assumed, shown by the yellow circles. A gene is a high-confidence target of a *trans*-regulatory factor if it contains the corresponding motif in its 3' UTR, and if its expression pattern across previously published expression datasets [33-39] significantly correlates with the expression patterns of other transcripts that contain that motif. Target genes are shown by blank circle, and the arrows demonstrate regulatory relationships.

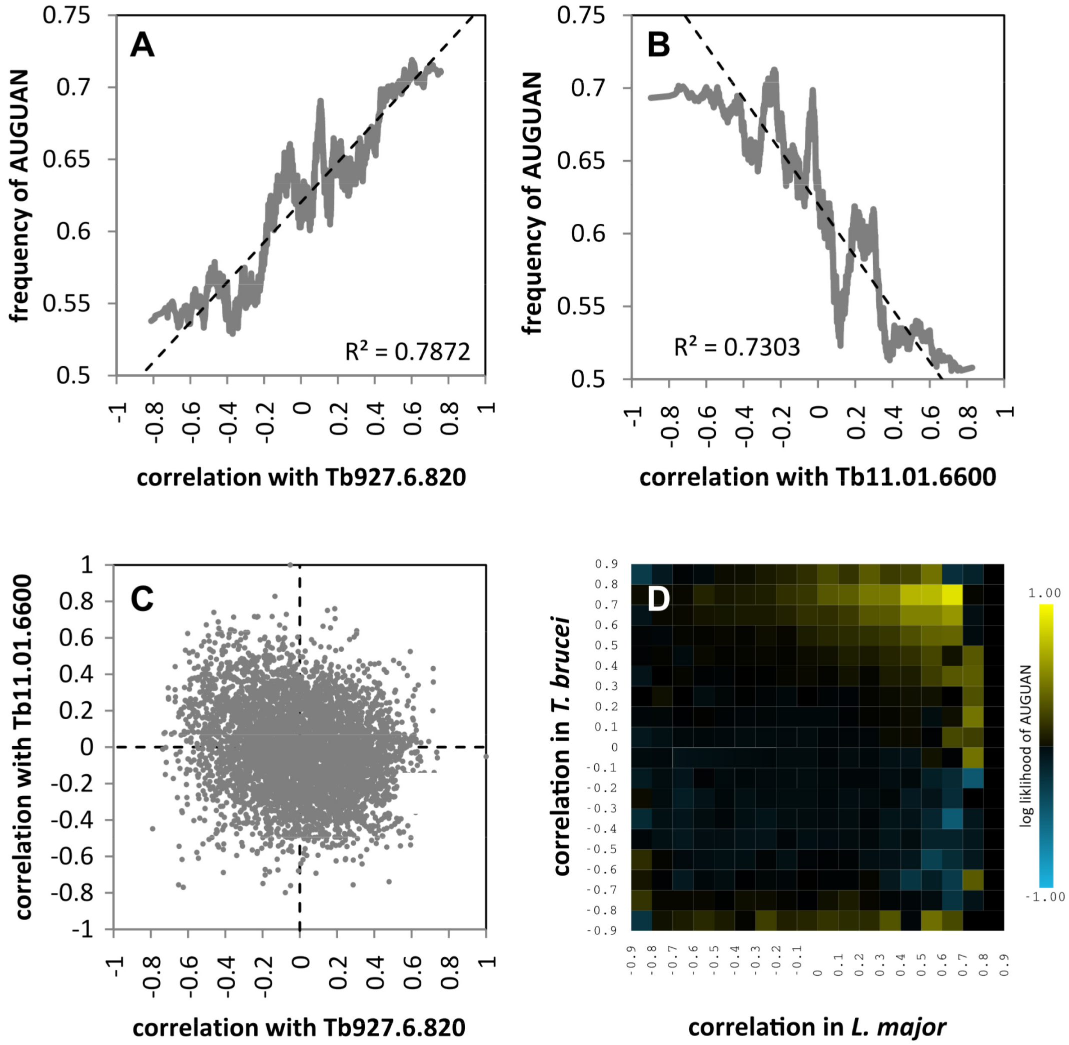


Figure 9-5. AUGUA as the potential binding sequence of PUF4 and PUF7 – The frequency of AUGUAN (Ptm17) is **(A)** significantly higher in transcripts that are positively correlated with PUF4 (Tb927.6.820, Mann-Whitney U z-score=7.12) and **(B)** in transcripts that are negatively correlated with PUF7 (Tb11.01.6600, Mann-Whitney U z-score=-8.46) across previously published expression datasets [33-39]. However, **(C)** the binding sites of PUF4 and PUF7 potentially contain additional nucleotides that confer specificity, as transcripts that are positively correlated with PUF4 are not necessarily the same as transcripts that are negatively correlated with PUF7. **(D)** AUGUAN has conserved its regulatory role across trypanosomatids including both the *Trypanosoma* genus and *Leismania* genus, as genes that are highly correlated in both *T. brucei* and *L. major* are more likely to have conserved their regulatory neighborhood. Two genes are assumed to have conserved their regulatory neighborhood if both of them contain the

AUGUAN motif in *T. brucei* as well as in *L. major*. To calculate the likelihood values, the set of such gene pairs were compared to a background set of gene pairs that are regulatory neighbors (i.e. both contain AUGUAN) in one organism, but have a broken regulatory neighborhood in the other organsm (i.e. only one of them contains AUGUAN).

9.2.4 A potential PUF-binding motif with a conserved regulatory role in trypanosomatids

As mentioned earlier, Ptrm17 contains the core binding sequence of PUF family of RNA-binding proteins, suggesting a major role of PUF family proteins in regulating the transcript stability and abundance in *T. brucei*. Interestingly, we found that the level of mRNAs containing this motif in their 3' UTRs shows very strong correlation or anti-correlation with the expression level of several PUF family proteins in *T. brucei*. Most notably, this motif is highly enriched among mRNAs that are co-expressed with PUF4 (Tb927.6.820) and among mRNAs that show strong anti-correlation with PUF7 (Tb11.01.6600) (Figure 9-5A,B), suggesting that these proteins bind to AUGUA-containing mRNAs and stabilize or destabilize them, respectively. However, the exact binding specificity and, therefore, target mRNAs of these two proteins may not be the same, as transcripts that are co-expressed with PUF4 are not necessarily the same as transcripts that are anti-correlated with PUF7 (Figure 9-5C). We also found that genes that have conserved instances of the AUGUA motif also have conserved expression patterns between *T. brucei* and *L. major* (Figure 9-5D), indicating that the regulatory function of this motif is highly conserved among trypanosomatids.

9.2.5 An AU-rich element (ARE) with a central role in regulating mRNA stability

Ptrm1970 is an AU-rich element (ARE) that contains the AUUUA sequence, which has been long known to be involved in regulation of mRNA stability in several eukaryotes [REF]. Different ARE-binding proteins have been characterized with a wide range of effects on target stability. While some ARE-binding proteins such as tristetraprolin (TTP), butyrate response factor 1 (BRF1) and AU-rich binding factor 1 (AUF1) destabilize their target transcripts [328], proteins of the ELAV-like family bind to AREs and stabilize the mRNA [329], mainly by protecting deadenylated transcripts against degradative enzymes [330]. Congruent with the protecting function of ELAV-like proteins, and based on analysis of previously reported microarray data [33], we found that Ptrm1970-containing transcripts are protected from degradation in *T. brucei* cells that over-express the DEAD-box RNA helicase DHH1 (Figure 9-2). DHH1 is required for efficient decapping of deadenylated mRNAs, which is an essential step in deadenylation-dependent decay pathway [331]. Also, we found that transcripts that have at least one instance of Ptrm1970 are over-represented in poly(A)$^+$ mRNA content of *T. brucei* cells that over-express poly(A)-specific ribonuclease 1 (PARN-1) [37], suggesting that Ptrm1970-containing transcripts are also protected against deadenylation activity of this enzyme (Figure 9-2). This is in line with previous reports showing that ELAV-like proteins can simultaneously bind to the ARE and poly(A) tail [332], and thus possibly protect the poly(A) tail of ARE-containing transcripts.

These observations suggest the presence of homologs of ELAV-like proteins in *T. brucei*, with a central role in regulation of mRNA stability via interaction with Ptm1970 and protection of the mRNA against deadenylation and/or deadenylation-dependent decay. It has been previously shown that the expression of human HuR, a member of the ELAV-like protein family, in *T. brucei* results in stabilization of several ARE-containing mRNAs [333]. However, the counterparts of HuR or any other ELAV-like proteins have not been characterized in *T. brucei*. Using PSI-BLAST, we found three potential remote homologs of ELAV-like proteins in *T. brucei*, Tb927.8.6650, Tb927.3.2930 and Tb927.7.5380. Intriguingly, we found that all three proteins are strongly co-regulated with transcripts that contain at least one instance of Ptm1970 (Figure 9-6A-C), supporting their function as ELAV-like proteins that bind and stabilize AREs.

As mentioned earlier, while ELAV-like proteins stabilize ARE-containing transcripts, other ARE-binding proteins may have an opposite effect. Exosome is a protein complex that is responsible for degradation of a wide variety of transcripts, and has been shown to directly interact with AREs and degrade ARE-containing transcripts [334, 335]. In search of proteins whose expression pattern is anti-correlated with Ptm1970, we found that the most significant anti-correlation belonged to the exosome subunit RRP45 (Tb927.6.670, Figure 9-6D). This protein has been previously suggested to have a role in initiation of rapid degradation of the very unstable mRNAs in *T. brucei* [336]. Our analysis suggests a central role of exosomes in regulation of mRNA stability in *T. brucei*, and proposes the exosome as the rate-limiting factor in ARE-mediated decay (AMD) in this organism.

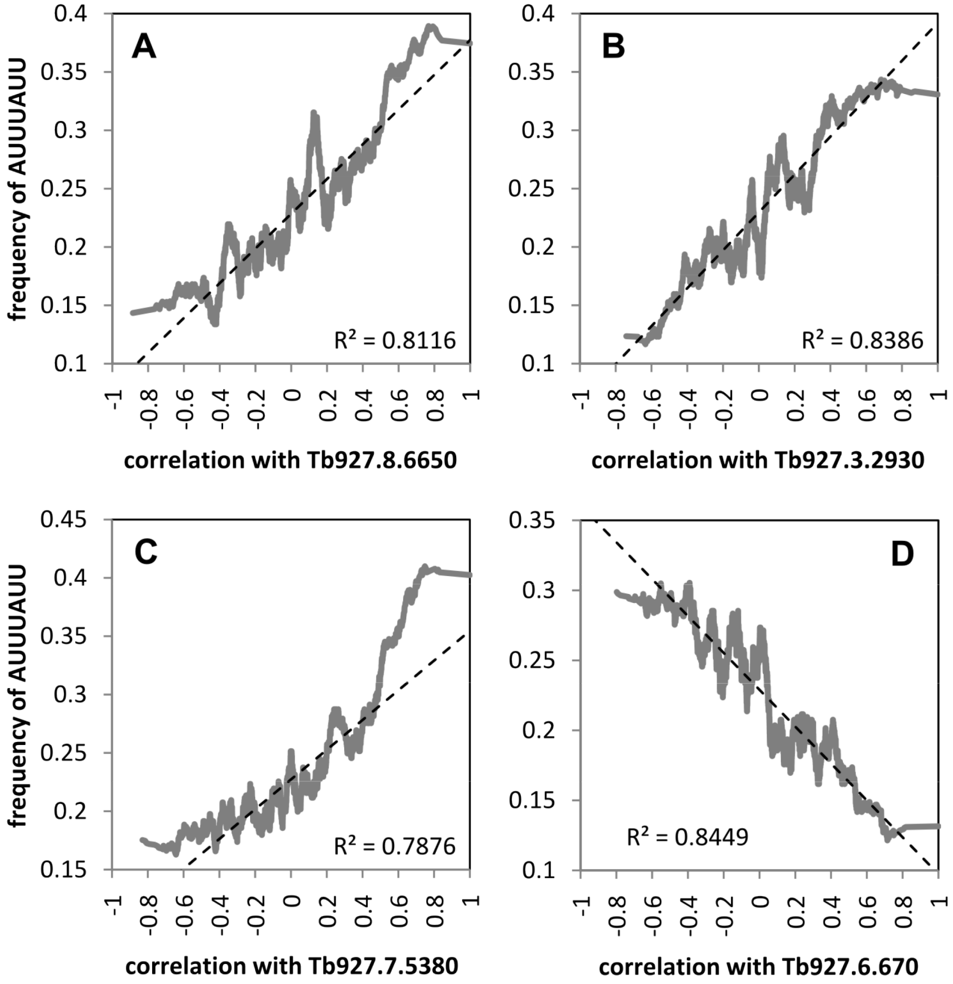


Figure 9-6. AU-rich elements are potentially regulated in *T. brucei* by three novel ELAV-like proteins as well as by exosome – (A-C) Three novel ELAV-like proteins that are identified using PSI-BLAST show significant positive correlation with transcripts that contain the AU-rich motif AUUUAUU (Ptm1970). This is shown here by plotting the frequency of AUUUAUU-containing transcripts against their correlation with each of the ELAV-like proteins across previously published *T. brucei* expression datasets [33-39]. **(D)** Transcripts whose expression patterns are negatively correlated with exosome complex exonuclease RRP45

(Tb927.6.670) are more likely to contain AUUUAUU, suggesting destabilizing effect of exosome on a large number of AUUUAUU-containing transcripts.

9.2.6 The GRN of T. brucei is responsive to external stimuli and internal perturbations

Early studies of *T. brucei* transcriptome suggested limited responsiveness of its GRN to external and internal perturbations within the same life stage [208]. Consequently, most studies of *T. brucei* transcriptome have focused on developmental gene regulation across different life stages of this parasite. Here, we examined the responsiveness of *T. brucei* transcriptome within the PF life stage by perturbing specific biological processes or imposing altered environmental conditions on the parasite. Specifically, we targeted mitochondrial DNA replication, protein synthesis, calcium ion transport, and cell cycle, and created environmental stress conditions using several chemical compounds. Microarray analysis revealed widespread remodeling of *T. brucei* transcriptome in response to these perturbations (Supplementary Figure 9-3). Interestingly, we found that transcripts carrying our predicted motifs show specific and coordinated responses to the perturbations (Figure 9-7), suggesting critical roles for several of these motifs in sensing and adapting to stress conditions as well as regulating biological processes. For example, transcripts that contain Ptm1970 are up-regulated when mitochondrial DNA replication is perturbed using ethidium bromide, and are down-regulated when calcium ion transport is blocked by verapamil and also when the growth medium is acidified by HCl. The motifs that responded to our set of chemically imposed perturbations also included five potential regulatory elements that could not be validated using previously available microarray data. For example, the linear motif Ptm25 (UYCGNGA) is specifically up-regulated in acidic conditions, and the structured Ptm2447 (NNGANCCAYNN) is specifically down-regulated when protein synthesis is inhibited by hygromycin (Figure 9-7). These experiments show that available microarray data of *T. brucei* are not representative of the complexity of its GRN, and suggest that the transcriptome remodeling should be examined in many different conditions and cell states in *T. brucei* in order to comprehensively characterize the regulatory programs of this organism and to understand the mechanisms that help this parasite adapt and respond to the environment.

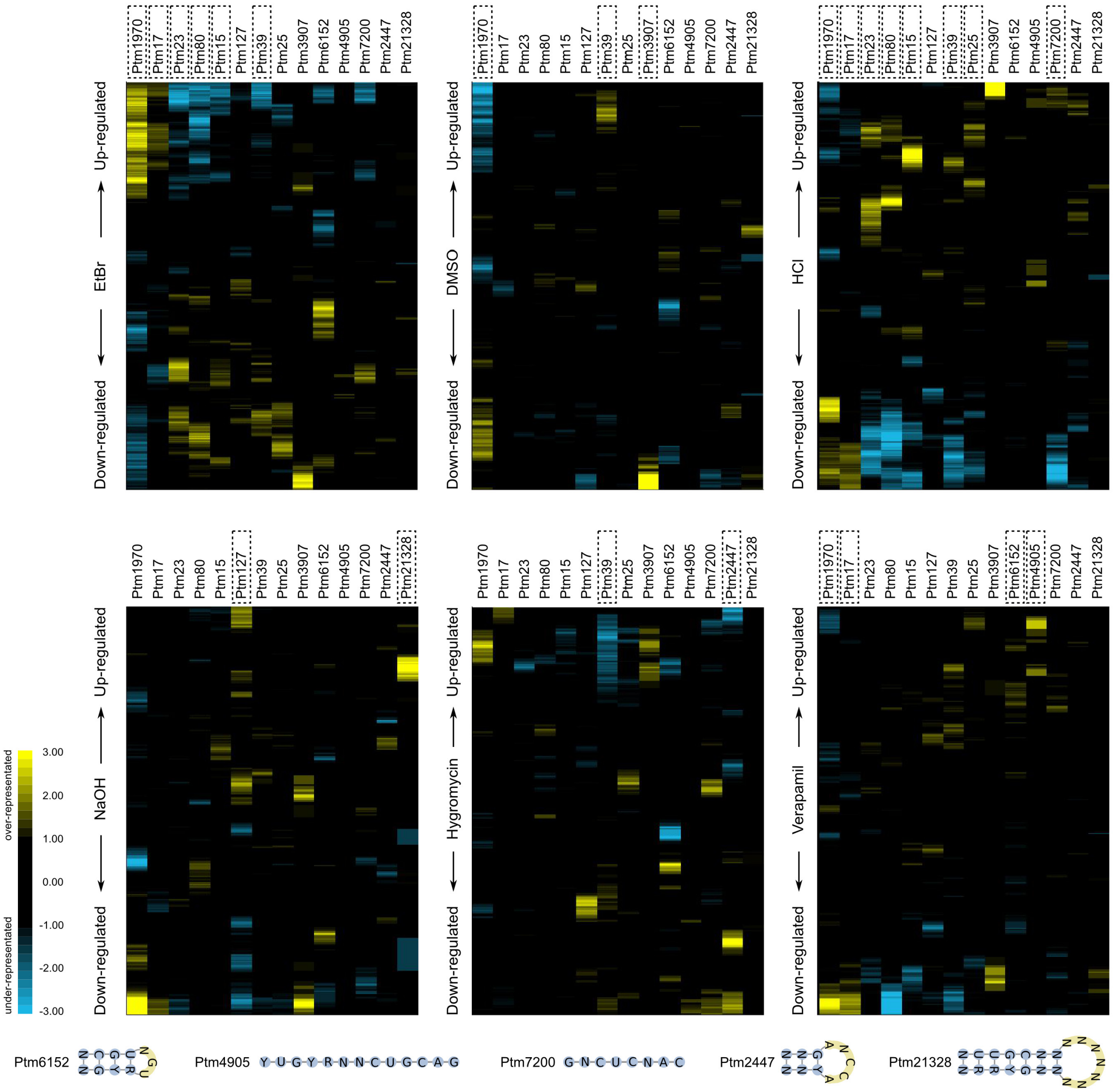


Figure 9-7. The GRN of *T. brucei* is responsive to environmental changes and stress conditions – Treatment of PF *T. brucei* cells with different chemicals results in significant up-/down-regulation of several conserved RNA-motifs. In each panel, genes that are up-regulated are shown on the top, and down-regulated genes on the bottom. Yellow indicates significant over-representation of motifs among genes with similar expression changes, and blue indicates significant under-representation of motifs. Over-/under-representation scores are calculated as the logarithm base 10 of cumulative *p*-values based on the Poisson

distribution, shown here by the yellow-black-blue color gradient. Motifs that are significantly up-regulated or down-regulated in each experiment (|Mann-Whitney U z-score| >3.62) are highlighted by the boxes. Five novel motifs were validated in this analysis, shown at the bottom. Ethidum bromide (EtBr) disrupts mitochondrial DNA replication and biogenesis in *T. brucei* [337]; DMSO has a wide range of effects on cell permeability and molecular interactions; hygromycin inhibits polypeptide synthesis in eukaryotes and prokaryotes [149]; and verapamil blocks calcium channels [338].

9.3 Conclusions

The statistical framework that we have introduced here provides a robust means for identification of regulatory programs that are conserved across multiple species. Most importantly, it provides an alignment-free method for identification of conserved *cis*-acting post-transcriptional regulatory motifs that contain sequence as well as structural information. This alignment-free framework allows identification of regulatory programs in genomes whose regulatory sequences have diverged extensively. Furthermore, in contrast to previous hypergeometric-based frameworks [320], our approach benefits from simultaneous analysis of multiple genomes, which leads to identification of motifs that are conserved in a large number of organisms.

We applied our framework to the genomes of trypanosomatids, a group of parasites with major health implications, in which regulatory mechanisms are poorly understood. Trypanosomatids provided a suitable benchmark for testing our framework, as gene regulation in these organisms is primarily post-transcriptional. Also, regulatory regions are poorly conserved in these organisms, meaning that conventional alignment-based approaches have limited applicability. Our approach was able to capture a large number of non-redundant *cis*-regulatory motifs, and we were able to validate 35 motifs using available expression data as well expression data obtained from targeted perturbation studies (Supplementary Table 9-2). Our analysis provides the first global picture of post-transcriptional regulatory programs in trypanosomatids, and identifies major regulatory roles for several new candidate *cis*- and *trans*-regulatory elements, including previously unidentified ELAV-like proteins. Further characterization of these candidate regulatory elements will not only lead into a better understanding of the biology of these parasites and the diseases they create, but also may provide new targets for chemical therapeutics that affect and disrupt conserved key regulatory functions in these organisms.

9.4 Methods

9.4.1 Trypanosomatid sequences and orthologs

Sequences of 3' untranslated regions (3' UTRs) for *Trypanosoma brucei*, *T. cruzi*, *T. vivax*, *T. congolense*, *Leishmania major*, *L. infantum*, *L. braziliensis*, and *L. mexicana* were downloaded from TriTrypDB v2.5 [339]. We defined the 3' UTR as the 1000-nucleotide region downstream of the stop codon, given that most 3' UTR motifs are known to reside within this region. Although in some cases this region may also contain a part of the downstream coding sequence, it is presumed that such contaminating sequences do not have an effect on the analysis.

Orthologous genes were identified based on OrthoMCL v4 [340]. Gene identifiers were converted to the most recent versions based on the list of gene aliases provided by TriTrypDB. Ortholog groups that contained several paralogs from the same organism were trimmed by randomly selecting one of the paralogs for each organism.

9.4.2 Identification of linear and structural motifs with signifincant network-level conservation

Supplementary Figure 9-1 schematically shows the statistical framework for identification of network-level conservation across multiple species, which is implemented in the software COSMOS (http://webpages.mcgill.ca/staff/Group2/rsalav/web/Software/COSMOS/index.htm). Consider *N* ortholog groups across *M* species with their associated regulatory sequences. We call an ortholog group a "keeper" of a motif if all *M* sequences that belong to this ortholog group have at least one instance of that motif. The number of these "keepers" determines the network-level conservation of the motif. COSMOS calculates the probability that a random distribution of the motif results in the observed number of keepers; small probability values indicate high conservation. The details of the calculations are as follows.

Given a particular motif and its instances in the regulatory regions of N genes in M species, the probability of occurrence of this motif in each species i is calculated as:

$$p_i = \frac{\sum_{1 \le j \le N} B\left(S_{i,j}\right)}{N},$$

where $B(S_{i,j})$ is 1 if the sequence that belongs to species i in the jth ortholog group has at least one instance of the motif; otherwise $B(S_{i,j})$ is 0. Thus, the probability of an ortholog group being a "keeper" under the null hypothesis of random distribution is:

$$p_{keeper} = \prod_{1 \le i \le M} p_i$$

The probability of observing at least n keepers is then calculated based on the binomial distribution:

$$p(K \ge n) = \sum_{n \le k \le N} f\left(k; N, p_{keeper}\right),$$

where f is the probability mass function of the binomial distribution, calculated as:

$$f(k; N, p) = \begin{pmatrix} N \\ k \end{pmatrix} p^k \left(1-p\right)^{N-k}$$

COSMOS uses $p(K \ge n)$ as the conservation score for a motif, in which n is the number of keepers of that motif.

In this work, we considered ~4.7×10^6 linear and structural motifs, including all possible linear motifs with a maximum length of 7 nt and maximum number of 6 non-degenerate bases, and all possible stem-loop motifs with a maximum stem length of 8 bp, loop length of 3-7 nt, and maximum number of 6 non-degenerate bases. COMOS calculated the conservation scores as well as the false discovery rate-adjusted p-values (q-values). Motifs with q-values ≤0.01 were retained.

9.4.3 Identification and removal of redundant motifs

Consider the more conserved motif i and the less conserved motif j in the list of motifs that are sorted by ascending order of their conservation p-value ($1 \le i < j$). Motif j is redundant given motif i if the instances of motif j in database D significantly overlap the instances of motif i. The overlap significance is calculated using the hypergeometric distribution as:

$$p_{j|i} = f\left(o_{i,j}; N_j, O_i, n_j\right),$$

where N_j is the total number of sliding windows of length l_j in database D (l_j is the length of motif j), O_i is the number of sliding windows of length l_j that overlap at least one nucleotide of at least one instance of motif i, n_j is the total number of instances of motif j, $o_{i,j}$ is the number of instances of motif j that overlap at least one nucleotide of at least one instance of motif i, and f is the probability mass function of the hypergeometric distribution. Motif j is considered redundant if there is at least one motif i ($i<j$) so that $p_{j|i} \le \frac{0.01}{j-1}$. The denominator j is for Bonferroni correction of p-value, as each motif j is compared to j-1 motifs that are better conserved. In this study, we used the 3’ UTRs of *T. brucei* genes as database D in order to identify redundant motifs.

9.4.4 Processing and merging available expression data

T. brucei microarray and RNA-seq data were obtained from multiple sources [33-39], and an expression compendium was compiled as follows. For E-MEXP-2025 and E-MEXP-2026 [33], the log ratio of induced vs. non-induced expression was calculated. For GSE18065 [36], the reported values were first averaged separately for BF as well as for PF. Then, the average of BF and PF was calculated for each gene, resulting in a single average measurement for each gene. The log ratio of average BF to the overall average measurement and also the log ratio of average PF to the overall average measurement were then calculated. For GSE20593 [37], all reported log ratios from biological replicates were averaged. For GSE22571 [38], the following experiments were averaged, resulting in a single measurement for each gene: GSM560209, GSM560212, GSM560213, GSM560214. Then, the log ratio of each experiment to this overall average measurement was calculated. Also, the log ratio of GSM560208:GSM560207 and the log ratio of GSM560211:GSM560210 were calculated. For GSE24275 [39], the total log ratio was used. The rest of the data were obtained from the supplementary information of the corresponding articles. Gene identifiers were converted to the most recent version according to TriTrypDB, and the datasets were merged into a single expression compendium. Genes that were present in only a subset of the datasets were included in the expression compendium; thus, the compendium contains missing values. Each experiment (column) was then normalized to have an average of zero and standard deviation of one across different genes, as described before [5]. It should be noted that this normalization does not affect single-array Mann-Whitney U-based analysis as described later, but is important for calculation of Pearson correlation coefficients across multiple experiments.

L. major microarray data were obtained from three previous publications [146, 165, 341]. The following sets of experiments were averaged since they were biological replicates: [GSM98805, GSM98806, GSM98870]; [GSM99790, GSM99791, GSM99792]; [GSM99795, GSM99796, GSM99797]; [GSM99798, GSM99799]; [GSM251641, GSM251642, GSM251643, GSM251644]; [GSM251645, GSM251646, GSM251647, GSM251648]; [GSM291427, GSM291428, GSM291429, GSM291430]. Datasets were merged and normalized as above.

9.4.5 Chemical treatment of PF T. brucei

PF *T. brucei* cells were treated with different chemicals and drugs in order to perturb specific biological processes or create environmental stress conditions. Wild-type PF *T. brucei* cell line IsTat 1.7A [163] was grown in SDM-79 medium in 26 °C, while the cell count was kept between 1×10^{7} and 3×10^{7} cells/ml. The cells were treated with either 2.65 µg/ml ethidium bromide, 3.1% (v/v) DMSO, 0.31% (v/v) HCl, 12.5 mM NaOH, 1.13 µg/ml hygromycin, 1.9 µM verapamil, 1.13 µg/ml G418, 130 nM pentamidine, 2.5×10^{-3} % (v/v) Triton-X, 31 ng/ml of phleomycin, or 7.8 mM imidazole. The concentrations were chosen based on the EC_{50} values of these chemicals for inhibition of *T. brucei* growth, as determined by growing *T. brucei* cells in the presence of different concentrations of chemicals. We chose the EC_{50} in order to ensure that the target biological process of each chemical is affected at the selected concentration.

9.4.6 Microarray analysis of chemical perturbations

T. brucei cells were collected 48 h after treatment, and total RNA was extracted using TRIzol Reagent (Invitrogen) and was further purified using RNeasy Mini Kit (Qiagen) as per manufacturers' instructions. RNA quality was examined using Agilent 2100 Bioanalyzer prior to cDNA preparation. 25 µg RNA was incubated with 9 µg oligo(dT) primer ($dT_{23}VN$, where V is a mixture of A, C and G, and N is any nucleotide) at 70 °C for 10 min, and Cy5-labeled cDNA was synthesized using Superscript III Reverse Transcriptase (Invitrogen) in the presence of 10mM DTT, 0.5 mM of each of dATP, dGTP and dTTP, 0.05 mM dCTP, and 0.05 mM of Cy5-dCTP as per manufacturer's instructions. Control cDNA from untreated PF *T. brucei* cells was prepared similarly using Cy3-dCTP. RNA was hydrolyzed by RNase A and RNase H, and cDNA was cleaned up using Qiagen PCR purification kit. Equal amounts of Cy3/Cy5-labeled cDNA were mixed and hybridized to version 4 of *T. brucei* microarrays from Pathogen Functional Genomics Resource Center as described before [164]. Microarrays were scanned using ScanArray Express (PerkinElmer) and the acquired images were quantified using ScanArray Express software with lowess normalization. The value for each probe was set to the binary logarithm of treated/control (Cy5/Cy3) median signal ratio, and different probes of each gene were averaged. Results of different chemical treatments were merged to obtain a matrix in which each row represented a gene and each column represented an experiment. Each column was then normalized to have an average of 0 and standard deviation of 1. Then, each row was normalized to have an average of 0. The latter normalization was aimed to neutralize the expression changes that represented a general response to stress and cell death, retaining only the expression changes that represented the specific response of the cells to the corresponding chemical treatment. Gene identifiers were converted to the most recent TriTrypDB version.

9.4.7 Identification of up- or down-regulated motifs

In order to identify motifs whose corresponding transcripts were significantly up- or down-regulated in previously published expression profiles or in our microarray data, we used the standard Mann-Whitney U test. For each motif, we compared the distribution of values for motif-containing transcripts to the distribution of values for transcripts that lacked that motif. For each microarray/RNA-seq experiment, all transcripts were sorted in the descending order of their corresponding values, and for each motif the sum of ranks of the transcripts that contained at least one instance was calculated. The z-score was then calculated as:

$$z=\frac{m_R-R}{\sigma_R},$$

where R is the sum of ranks, m_R is the average expected R calculated as $[n(n+1) +n\times \acute{n}]/2$ (where n is the number of transcripts that have the motif and $\acute{n}$ is the number of transcripts that do not have any instance of the motif), and σ_R is the standard deviation of R, calculated as:

$$\sigma_R=\sqrt{\frac{n\cdot n'\left(n+n'+1\right)}{12}}$$

Positive z-scores indicate up-regulation, and negative z-scores indicate down-regulation. False discovery rate (FDR) at different z-score cutoffs was determined by performing the same analysis on 100 randomly shuffled motif occurrence profiles, setting the cutoff at FDR ≤0.1.

9.4.8 Identification of motifs with local enrichment in the expression hyperspace

In order to identify the motifs that occurred in co-regulated transcripts across the previously published microarray/RNA-seq datasets [33-39], we searched for regions in the expression hyperspace where motifs were significantly enriched. Each point in this hyperspace was expressed as a vector of size 22, representing the 22 measurements in the datasets from previous publications. For each point in the hyperspace, genes were first sorted by the descending order of their Pearson correlation with the vector that corresponded to that point, and the local enrichment of each motif was measured using Mann-Whitney U test. Significant positive z-scores were determined by shuffling the motif occurrence profiles across the transcripts and recalculating the z-scores as described in the previous section (FDR≤0.1). Using this method, we examined motif enrichment around 8035 points in the expression hyperspace, each point corresponding to a *T. brucei* gene. By adapting this approach, we limited our search to regions in the hyperspace that were populated by genes, rather than deserted regions. This approach also identified genes whose expression profiles were highly correlated with certain motifs; this information was used to construct the high-confidence GRN of *T. brucei*.

9.5 Supplementary Tables

Supplementary Table 9-1. All of the 388 non-redundant motifs identified in this study based on network-level conservation in the genus of *Trypanosoma* –For more details on motifs that were supported by microarray data, see **Supplementary Table 9-2**.

Motif	Sequence	Structure	Conservation *p*-value (sequence and structure)	Conservation *p*-value (sequence only)
Ptm1	CAUAGAN	....	2.12E-29	2.12E-29
Ptm13	UNAUGGA	(.....)	1.43E-19	1.43E-19
Ptm15	UYGANGA	((...))	1.64E-18	1.64E-18
Ptm17	AUGUAN	......	3.65E-18	3.65E-18
Ptm23	UYGCNGA	((...))	1.06E-17	1.06E-17
Ptm25	UYCGNGA	((...))	3.95E-17	3.95E-17
Ptm32	AUGGGNRU	((....))	2.03E-16	2.03E-16
Ptm39	GCNCCNNNNGY	((.......))	5.02E-16	5.02E-16
Ptm53	NNUCAGCUNNN	((.......))	4.05E-15	0.000291551
Ptm60	UUYGNNNNCACGAA	((((......))))	1.83E-14	1.83E-14
Ptm69	GANUGGNNNNNRNUY	((((.......))))	4.97E-14	1.59E-08
Ptm80	NNCAGGANN	(.......)	9.42E-14	9.54E-07
Ptm121	AUGUCNNNNNRYRU	((((......))))	8.60E-13	8.60E-13
Ptm127	CNAAGGNNG	(.......)	1.24E-12	1.24E-12
Ptm131	NNUYNACAAGANN	((((.....))))	1.42E-12	1.29E-06
Ptm176	UGGCCNNNYYR	(((.....)))	4.81E-12	4.81E-12
Ptm185	NRYUNNNCAAGUN	((((.....))))	6.86E-12	1.63E-07
Ptm186	NYNCGCAAGN	((......))	6.99E-12	1.81E-08
Ptm198	NGNUUNNNNNANAANCN	(((((.......)))))	1.09E-11	2.08E-05
Ptm212	GGUNCANRYY	(((....)))	1.38E-11	1.38E-11
Ptm234	NAGGCNAUN	((.....))	2.12E-11	6.33E-09
Ptm264	NNUNNAGAACANN	(((.......)))	3.54E-11	9.23E-07
Ptm298	NYAAGAAGN	((.....))	5.59E-11	1.12E-10
Ptm306	NNUYGUNNACACGANN	((((((....))))))	6.21E-11	1.86E-08
Ptm307	GCUCAGNNNNNNYUGRGY	(((((((....)))))))	6.21E-11	2.48E-10
Ptm332	NNGRYNNNNNGUGUCNN	(((((.......)))))	8.46E-11	0.000107675
Ptm360	GGUNNANUACC	(((.....)))	1.07E-10	1.07E-10
Ptm384	NNNNRGRNNAAGUCUNNNN	(((((((.....)))))))	1.24E-10	8.35E-06
Ptm429	GNCUGGAC	(......)	2.07E-10	2.07E-10
Ptm444	NNGGGNNNNAAACCCNN	(((((.......)))))	2.48E-10	1.39E-07
Ptm445	NNGUAUCCNRURYNN	((((((...))))))	2.48E-10	3.13E-07

Ptm446	GYGRGUNNNACUCGC	((((((...))))))	2.48E-10	2.48E-10
Ptm491	NNNNCAAGGANUGNNNN	((((((.....))))))	3.11E-10	0.000201617
Ptm495	NAGNUNCNNNNNRNYUN	(((((.......)))))	3.20E-10	0.00403603
Ptm505	GGYNNNNGAUGCC	(((.......)))	3.46E-10	3.46E-10
Ptm509	NNNRNNAUNCUNNN	((((......))))	3.53E-10	0.00254345
Ptm519	NUAUGAGNNNNYUYRURN	(((((((....)))))))	3.73E-10	2.98E-09
Ptm541	NNYNNUACGAGNN	(((.......)))	4.37E-10	0.00119644
Ptm580	ACUGGARGU	(((...)))	5.46E-10	5.46E-10
Ptm591	NRUNNNNCGCAUN	(((.......)))	6.04E-10	2.72E-10
Ptm594	GUGRNNUUCAC	((((...))))	6.08E-10	6.08E-10
Ptm636	GGNCAGCC	((....))	7.87E-10	7.87E-10
Ptm666	NCGGNUCNNNYYGN	((((......))))	9.85E-10	1.58E-06
Ptm670	NGGAGCUNNNNNGYUYYN	((((((......))))))	9.94E-10	6.21E-08
Ptm728	YUNAANCAG	((.....))	1.23E-09	1.23E-09
Ptm765	YGYUYUNNNNAGAGCG	((((((....))))))	1.49E-09	1.49E-09
Ptm776	NNNNNGRNCCNCUCNNNNN	(((((((.....)))))))	1.49E-09	5.47E-06
Ptm783	YNUACCAG	(......)	1.62E-09	1.62E-09
Ptm831	ANNUNGNNNNNNNYNRNNU	((((((.......))))))	1.98E-09	1.37E-07
Ptm836	NNNNYGRNNCNCUCGNNNN	(((((((.....)))))))	1.99E-09	0.00409599
Ptm866	GNUYUUNNNNAAGANC	((((((....))))))	2.24E-09	4.17E-08
Ptm906	CCGUCGNNGRYGG	(((((...)))))	2.48E-09	2.48E-09
Ptm918	RGYRGNNNCUGCU	(((((...)))))	2.60E-09	2.60E-09
Ptm926	NNAUNAACNNNRUNN	((((.......))))	2.68E-09	6.93E-07
Ptm947	NNCCGCAUYGGNN	(((((...)))))	2.98E-09	1.21E-06
Ptm986	GCGCCGNNNNYGGYGY	((((((....))))))	3.35E-09	3.35E-09
Ptm998	RNYUNNNNCGAGNU	((((......))))	3.44E-09	2.13E-06
Ptm1003	UGYYNNUGGCA	((((...))))	3.50E-09	3.50E-09
Ptm1016	CCAGGCNUGG	(((....)))	3.73E-09	3.73E-09
Ptm1018	NNNNGNNUCAAACNNNN	(((((.......)))))	3.73E-09	0.00424737
Ptm1019	NGUACGGNNNNNYGURYN	((((((......))))))	3.73E-09	6.26E-08
Ptm1032	NNNNNNACCUNGNNNNNNN	((((((.......))))))	3.98E-09	0.000266377
Ptm1043	NNAGCANAUNN	(((.....)))	4.11E-09	0.00475076
Ptm1081	NYYRRNNNNNAUUGGN	(((((......)))))	4.51E-09	1.36E-08
Ptm1091	NNGAUGCNCNN	(((.....)))	4.70E-09	8.27E-10
Ptm1139	NYYUNAAGGN	(((....)))	5.42E-09	2.45E-05
Ptm1188	NCGGGCUNNNNNYYYGN	(((((.......)))))	5.96E-09	5.22E-08
Ptm1192	UYNUGRNNNNNUCANGA	((((((.....))))))	5.96E-09	1.88E-07
Ptm1293	CGANCUNNNNNNGNUYG	(((((.......)))))	7.45E-09	2.61E-07
Ptm1311	NNNNGUGGAUNYRYNNNN	(((((((....)))))))	7.83E-09	2.91E-07
Ptm1359	UYNGGNNNGCCNGA	(((((....)))))	8.94E-09	2.52E-07
Ptm1363	NNNAUGGUGNNYRUNNN	((((((.....))))))	8.94E-09	0.000576733

Ptm1369	NGRNYNGNNNNNNNCNGNUCN	(((((((.......)))))))	8.94E-09	0.000644473
Ptm1374	NUNUUNNCNAANAN	(((((....)))))	9.19E-09	0.000425128
Ptm1422	NNNNYGUNGAGNNNN	(((((.....)))))	1.02E-08	4.17E-06
Ptm1423	NNGANGCNNUYNN	(((((...)))))	1.02E-08	3.46E-10
Ptm1436	NNNNNNNGGCGANNNNNNNNN	(((((((.......)))))))	1.04E-08	0.0287297
Ptm1447	NYGNNNUNAACGN	(((.......)))	1.07E-08	0.000887285
Ptm1450	YYRRNNUUUGG	((((...))))	1.07E-08	1.07E-08
Ptm1455	NNYGNNAAGNCGNN	((((......))))	1.10E-08	0.00144562
Ptm1515	NNNYYGRNNNNCUCGGNNN	(((((((.....)))))))	1.19E-08	5.83E-05
Ptm1534	UGYYNNNNGGGCA	((((.....))))	1.25E-08	1.25E-08
Ptm1615	NGCCGUANNGGYN	((((.....))))	1.49E-08	1.05E-06
Ptm1617	CAUGGGNNNNYYYRUG	((((((....))))))	1.49E-08	1.49E-08
Ptm1618	CCNAUGNNNNNYRUNGG	((((((.....))))))	1.49E-08	1.41E-14
Ptm1663	NNUGGCACNNNNNGYYRNN	((((((.......))))))	1.68E-08	1.77E-05
Ptm1664	NNYGGNNNNNNAANCCGNN	((((((.......))))))	1.68E-08	0.000171728
Ptm1691	CCAUACNNNRUGG	((((.....))))	1.79E-08	1.79E-08
Ptm1694	NNNNGGNNNCANACCNNNN	((((((.......))))))	1.79E-08	0.00366707
Ptm1725	NNNNNGNUGGUGCNNNNN	((((((......))))))	1.86E-08	0.0104629
Ptm1727	URYGNGNNNNNNNCNCGUA	((((((.......))))))	1.86E-08	4.93E-07
Ptm1773	NYUURRNNNNNNNUUAAGN	((((((.......))))))	2.01E-08	1.97E-06
Ptm1810	NNRUUCNGAAUNN	(((((...)))))	2.17E-08	8.21E-05
Ptm1829	NCCGUUCNNNNYGGN	((((.......))))	2.24E-08	1.20E-06
Ptm1837	NNNNNGCUUCANNYNNNNN	((((((.......))))))	2.24E-08	0.175109
Ptm1882	NNGUNAAGAACNN	((((.....))))	2.48E-08	1.40E-09
Ptm1909	NNNNAGCNGGNGYUNNNN	(((((((....)))))))	2.61E-08	4.54E-06
Ptm1910	NNNNANCUNGNGNUNNNN	(((((((....)))))))	2.61E-08	2.24E-05
Ptm1934	NNNNGANGGCNNNNNUYNNNN	(((((((.......)))))))	2.68E-08	0.00657004
Ptm1950	GCNCAGNNNNNNYUGNGY	((((((......))))))	2.78E-08	3.49E-07
Ptm1953	GCNNAGNNNNNNYUNNGY	(((((((....)))))))	2.80E-08	0.00845982
Ptm1970	AUUUAUU	(.....)	2.86E-08	2.86E-08
Ptm2002	NNNCGUAAGGNNN	((((.....))))	2.98E-08	0.00226332
Ptm2004	NGUAUCCNNNRURYN	(((((.....)))))	2.98E-08	3.49E-07
Ptm2009	NCNCAGNNNNNNYUGNGN	(((((((....)))))))	2.98E-08	0.0012397
Ptm2010	NNYYRGYNNNNGCUGGNN	(((((((....)))))))	2.98E-08	2.11E-05
Ptm2239	NRGYNCGGGCUN	((((....))))	4.17E-08	1.78E-06
Ptm2240	NGAUAGGNNNRUYN	((((......))))	4.17E-08	1.14E-06
Ptm2279	AGUCGUNGRYU	((((...))))	4.47E-08	4.47E-08
Ptm2281	NNYGUNCACACGNN	(((((....)))))	4.47E-08	2.99E-05
Ptm2283	NGRYGNNNNGCCGUCN	(((((......)))))	4.47E-08	1.05E-06
Ptm2284	UGGAACNNNNNNUUYYR	(((((.......)))))	4.47E-08	4.47E-08
Ptm2335	GNGRGNNNNACUCNC	(((((.....)))))	4.77E-08	2.85E-06

Ptm2405	NNUNAANCCANN	(((......)))	5.29E-08	2.05E-05
Ptm2447	NNGANCCAYNN	(((.....)))	5.58E-08	0.194717
Ptm2454	GCNAGUNNNRYUNGY	((((((...))))))	5.59E-08	1.31E-06
Ptm2469	NUCAGNUNNNNYUGRN	((((((....))))))	5.74E-08	8.01E-05
Ptm2534	YGGYNNGCGCCG	((((....))))	6.21E-08	6.21E-08
Ptm2553	UYGNNUACGA	(((....)))	6.33E-08	6.33E-08
Ptm2597	YGRGUNNNACUCG	(((((...)))))	6.71E-08	6.71E-08
Ptm2600	NNNGGCUGANNYYNNN	(((((......)))))	6.71E-08	0.00566673
Ptm2604	UGCGGGNNNNNNNYYYGYR	((((((.......))))))	6.71E-08	6.71E-08
Ptm2698	URNNGGNNNNNNNCCNNUA	((((((.......))))))	7.61E-08	0.000169706
Ptm2720	GYGGNAACCGC	((((...))))	7.83E-08	7.83E-08
Ptm2751	GUUUGANNNNNNNUYRRRY	((((((.......))))))	8.05E-08	8.05E-08
Ptm2843	GYYNGNNNACNGGC	(((((....)))))	8.94E-08	7.21E-06
Ptm2863	GUANUCNNNNGRNURY	((((((....))))))	9.13E-08	7.38E-07
Ptm2871	CAAAAC	......	9.16E-08	9.16E-08
Ptm2933	NNGGNUACNYYNN	((((.....))))	1.01E-07	0.000316367
Ptm2967	ACAUGGNNNNNRUGU	((((.......))))	1.04E-07	1.04E-07
Ptm2968	NNNUNNGGGCAANNN	((((.......))))	1.04E-07	0.0102392
Ptm3033	GACUACNUY	((.....))	1.14E-07	1.14E-07
Ptm3049	GYYGNNNNNUUCGGC	((((.......))))	1.15E-07	1.15E-07
Ptm3050	NNNNUGANGUNNUYRNNNN	(((((((.....)))))))	1.15E-07	2.47E-05
Ptm3064	NNUCNGCUNNNNYNGRNN	((((((......))))))	1.17E-07	0.00544856
Ptm3065	YYRRUNNNNNNNNNAUUGG	(((((((.....)))))))	1.17E-07	6.69E-05
Ptm3078	CGAUGUNNNNNYRUYG	(((((......)))))	1.19E-07	1.19E-07
Ptm3079	NNRUYNNNNAANGAUNN	((((((.....))))))	1.19E-07	2.12E-05
Ptm3102	NNACNAACNNGUNN	(((((....)))))	1.22E-07	0.00835425
Ptm3108	NNNNCGGNAUNNYGNNNN	((((((......))))))	1.22E-07	0.00251532
Ptm3146	NGNUACARNYN	((((...))))	1.26E-07	2.00E-06
Ptm3150	NNAUAGUUNNNRUNN	((((.......))))	1.27E-07	0.000659621
Ptm3163	NNNAGCUGNNGYUNNN	((((((....))))))	1.28E-07	2.30E-05
Ptm3209	GRRYRNNNNNAUGUUC	(((((......)))))	1.34E-07	1.34E-07
Ptm3227	UGRNUNAUCA	(((....)))	1.36E-07	1.36E-07
Ptm3296	CGUACUNNNNURYG	((((......))))	1.43E-07	1.43E-07
Ptm3359	UGGAUGNNNNNYRUYYR	((((((.....))))))	1.52E-07	1.52E-07
Ptm3368	GNRUYNNNNAGAUNC	(((((.....)))))	1.53E-07	7.54E-06
Ptm3378	NUGUAUCNNNURYRN	(((((.....)))))	1.55E-07	4.07E-06
Ptm3391	NNCAUCGUNUGNN	((((.....))))	1.57E-07	0.000174799
Ptm3393	NNYGUCNAACGNN	(((((...)))))	1.57E-07	0.000232195
Ptm3489	NAUGANANNNNNNUYRUN	((((((......))))))	1.70E-07	0.000619992
Ptm3558	NGUUUUGNRRYN	((((....))))	1.82E-07	2.29E-06
Ptm3596	NNURNNCCCUUANN	((((......))))	1.86E-07	0.000238906

Ptm3708	AUCAAGNNNNUGRU	((((......))))	2.09E-07	2.09E-07
Ptm3709	YRUYGNNNNNGCGAUG	(((((......)))))	2.09E-07	2.09E-07
Ptm3762	NNAUUCAGNNN	((.......))	2.17E-07	0.000242812
Ptm3787	UGGYNNNNNGNGCCA	(((((.....)))))	2.23E-07	8.28E-06
Ptm3830	UGNCNUACA	((.....))	2.31E-07	2.31E-07
Ptm3880	NNNACUCCCNUNNN	((((......))))	2.39E-07	0.0247597
Ptm3907	YUUYRNNAUGAAG	(((((...)))))	2.46E-07	2.46E-07
Ptm3921	NNGGNNUANACCNN	((((......))))	2.48E-07	0.000635673
Ptm3926	NNNNUNNAACUNANNNN	(((((.......)))))	2.50E-07	0.056462
Ptm3981	CGNAUCNNNNNYG	(((.......)))	2.61E-07	8.76E-06
Ptm4080	NNYGUNNNNCCNACGNN	(((((.......)))))	2.78E-07	0.000447124
Ptm4119	CCGGUGNNNYYGG	((((.....))))	2.86E-07	2.86E-07
Ptm4197	UGYCNAGCA	(((...)))	3.03E-07	3.03E-07
Ptm4227	CGCGGUNYG	((.....))	3.11E-07	3.11E-07
Ptm4238	RUGGYNNNNNGCCAU	(((((.....)))))	3.13E-07	3.13E-07
Ptm4323	GCANGGNNNYYNUGY	((((((...))))))	3.34E-07	9.20E-06
Ptm4393	NNCUACCCNGNN	(((......)))	3.52E-07	0.000401341
Ptm4477	NGUNNNAGUUACN	(((.......)))	3.76E-07	1.78E-05
Ptm4479	UACGACNNYGUR	((((....))))	3.76E-07	3.76E-07
Ptm4485	NNNAGGACAUNNN	((((.....))))	3.76E-07	5.19E-05
Ptm4513	GAUUCUNNNRRUY	((((.....))))	3.88E-07	3.88E-07
Ptm4517	UGACNGA	(.....)	3.88E-07	3.88E-07
Ptm4559	RRRNNNGAGUUU	(((......)))	3.99E-07	3.99E-07
Ptm4625	NNGYNANGNGCNN	(((((...)))))	4.20E-07	1.59E-06
Ptm4628	CCCGUAGG	((....))	4.21E-07	4.21E-07
Ptm4688	NYUGNNNUCGCAGN	((((......))))	4.35E-07	9.12E-06
Ptm4715	NNNUNUGGAANANNN	((((((...))))))	4.43E-07	7.44E-05
Ptm4769	RUYUNNNNGGAGAU	((((......))))	4.59E-07	4.59E-07
Ptm4834	NGCGGNCNNNNYYGYN	((((((....))))))	4.78E-07	0.000323415
Ptm4889	NNAUACAANRUNN	((((.....))))	4.92E-07	0.00139358
Ptm4905	YUGYRNNCUGCAG	(((((...)))))	4.97E-07	4.97E-07
Ptm4909	NYUUNCACAAGN	((((....))))	5.01E-07	4.09E-05
Ptm4919	NNGGYNNGGNGCCNN	(((((.....)))))	5.03E-07	0.000881915
Ptm4923	NUUUGGCNNNNYYRRRN	((((((.....))))))	5.03E-07	7.97E-06
Ptm4928	NNNNNGCGGGCNNNN	((((.......))))	5.07E-07	0.111492
Ptm4983	NNNNGCUNGCNNGYNNNN	((((((......))))))	5.22E-07	0.000349604
Ptm5069	CAGGGNNNYUG	(((.....)))	5.53E-07	5.53E-07
Ptm5078	NNNNNNGNCCNCYNNNNNN	(((((((.....)))))))	5.57E-07	0.0226062
Ptm5103	NNNNGGCGANNNYYNNNN	((((((......))))))	5.64E-07	0.345663
Ptm5171	RUYYGNNNCGGAU	(((((...)))))	5.84E-07	5.84E-07
Ptm5227	NNNNGNNCCGNGCNNNN	(((((.......)))))	6.01E-07	0.0282279

Ptm5228	UCCAGANNNNGGR	(((.......)))	6.03E-07	6.03E-07
Ptm5254	GYYYNNNCUGGGC	((((.....))))	6.10E-07	6.10E-07
Ptm5292	CUUUGGNNYRRRG	(((((...)))))	6.20E-07	6.20E-07
Ptm5409	NGUNNNAAGCACN	(((.......)))	6.64E-07	4.50E-05
Ptm5416	NGCANGGNNNYNUGYN	((((((....))))))	6.68E-07	0.00033097
Ptm5562	GYGNGNNNNNCNCGC	((((((...))))))	7.24E-07	0.00113121
Ptm5649	NNNNNNAUGAUGNNNNN	(((((.......)))))	7.51E-07	1.51E-08
Ptm5728	NGCGCNNNNNNNGYGYN	(((((((...)))))))	7.83E-07	0.0217762
Ptm5805	NUYUGCAAGAN	((((...))))	8.14E-07	1.82E-05
Ptm5827	NNYUYUNNUNAGAGNN	((((((....))))))	8.29E-07	0.000224327
Ptm5914	NAUGCGNNNNNYRUN	((((.......))))	8.71E-07	0.000712759
Ptm5939	NNNNNGUUACANNNN	((((.......))))	8.79E-07	0.113111
Ptm5987	NGRRNUACUUCN	((((....))))	9.02E-07	4.28E-05
Ptm6030	GRGNCACCUC	(((....)))	9.22E-07	9.22E-07
Ptm6126	NGGUNNCNCACCN	((((.....))))	9.69E-07	5.03E-05
Ptm6152	NCGUNGURYGN	((((...))))	9.82E-07	2.96E-05
Ptm6200	GYYRNNNNNAUGGC	((((......))))	1.01E-06	1.01E-06
Ptm6217	UNGYRYNNNNGUGCNA	((((((....))))))	1.02E-06	2.23E-05
Ptm6261	NYUUYNNNNNAUGAAGN	(((((.......)))))	1.04E-06	1.70E-05
Ptm6302	NNGUCNGNNNNNYNGRYNN	(((((((.....)))))))	1.07E-06	0.0174916
Ptm6345	NCAUCAUNNNRUGN	((((......))))	1.09E-06	2.64E-05
Ptm6362	GGYRNGAUGCC	((((...))))	1.10E-06	1.10E-06
Ptm6465	NRGRNNNNNCANUCUN	(((((......)))))	1.14E-06	0.00259682
Ptm6488	YRUYRYNNNNNNNGUGAUG	((((((.......))))))	1.16E-06	1.16E-06
Ptm6602	GCAUGGNRUGY	((((...))))	1.22E-06	1.22E-06
Ptm6650	NNNGAAUUGNYNNN	((((......))))	1.25E-06	0.220239
Ptm6693	NNNNAGACNANNN	(((.......)))	1.27E-06	6.64E-05
Ptm6763	NYRUYNNNUGGAUGN	(((((.....)))))	1.31E-06	4.68E-05
Ptm6764	NAGNNCANNNNNNGNNYUN	((((((.......))))))	1.31E-06	0.0846544
Ptm6864	CAUGCANNNNNGYRUG	(((((......)))))	1.38E-06	1.38E-06
Ptm6870	NGUNNAANNACN	((((....))))	1.38E-06	0.0121621
Ptm6871	NGACCGCUYN	(((....)))	1.38E-06	0.000143047
Ptm6887	NNGGACCACNN	(((.....)))	1.39E-06	0.00276236
Ptm7029	GUGGGGNNNYYYYRY	((((((...))))))	1.47E-06	1.47E-06
Ptm7127	AGUCNGNNNNRYU	(((.......)))	1.53E-06	1.53E-06
Ptm7129	YRUGNNNGACAUG	((((.....))))	1.53E-06	1.53E-06
Ptm7142	NNGYUNNGGNAGCNN	(((((.....)))))	1.54E-06	0.00293888
Ptm7200	GNCUCNAC	(......)	1.57E-06	1.57E-06
Ptm7359	NNUNCCGUAANN	(((......)))	1.70E-06	0.00197858
Ptm7443	NNGGNNNGNGACCNN	((((.......))))	1.75E-06	4.54E-06
Ptm7492	GUGNNNNCAGCAC	(((.......)))	1.80E-06	1.80E-06

Ptm7495	NNYGNNUACACGNN	((((......))))	1.80E-06	0.00230365
Ptm7521	NGUNCGCCACN	(((.....)))	1.82E-06	9.80E-05
Ptm7549	NNNGNNNUGGGNCNNN	(((((......)))))	1.84E-06	0.000137082
Ptm7552	GUUYNNNNAGAAC	((((.....))))	1.85E-06	1.85E-06
Ptm7649	UAGGCGNNYUR	(((.....)))	1.92E-06	1.92E-06
Ptm7721	NNNNCAGUUNGNNNN	(((((.....)))))	1.97E-06	0.000326218
Ptm7809	NNNNYRNGANCUGNNNN	((((((.....))))))	2.03E-06	2.77E-05
Ptm7810	ANUGGUNNNNNNRYYRNU	(((((((....)))))))	2.03E-06	1.15E-07
Ptm7840	NRGURNNUNUACUN	(((((....)))))	2.06E-06	6.20E-05
Ptm7859	NNNNUGACANYRNNNN	((((((....))))))	2.08E-06	0.0126455
Ptm7914	RGRYNNNNNNNANGUCU	(((((.......)))))	2.12E-06	6.69E-05
Ptm7943	GGRNNNNCUCUCC	(((.......)))	2.15E-06	2.15E-06
Ptm7977	NCCUNGGNNNRGGN	((((......))))	2.17E-06	0.000105933
Ptm8019	NNGAUGCGNNUYNN	((((......))))	2.21E-06	0.00283139
Ptm8124	YGGNGGGCCG	(((....)))	2.30E-06	2.30E-06
Ptm8170	CUNAUUNNNNNNRUNRG	(((((.......)))))	2.35E-06	3.87E-05
Ptm8225	GRGNRRNNNUUNCUC	((((((...))))))	2.41E-06	4.08E-05
Ptm8241	CANCCANNNGNUG	((((.....))))	2.42E-06	0.000188702
Ptm8259	GRNUGCAUC	((.....))	2.43E-06	2.43E-06
Ptm8509	RRGYNNNNCAGCUU	((((......))))	2.66E-06	2.66E-06
Ptm8525	NGCAGCANUGYN	((((....))))	2.67E-06	0.000162666
Ptm8680	NNGGCNCCNNGYYNN	(((((.....)))))	2.82E-06	0.0039588
Ptm8756	NYRUNRNNUUNAUGN	((((((...))))))	2.89E-06	0.00778467
Ptm8941	CGAACNNNNUUYG	((((.....))))	3.07E-06	3.07E-06
Ptm8962	GUUAUUNNRRY	(((.....)))	3.10E-06	3.10E-06
Ptm9227	NAGNCUANNNNNYUN	((((.......))))	3.38E-06	1.04E-05
Ptm9289	NNACNGGUNNNGUNN	((((.......))))	3.45E-06	0.00815134
Ptm9336	NNANUUNGNNNNNRRNUNN	(((((((.....)))))))	3.51E-06	0.000608797
Ptm9505	GGNNCCGCCC	((......))	3.71E-06	3.71E-06
Ptm9518	RYYUYNNNNNNGAGGU	((((((....))))))	3.73E-06	0.000212367
Ptm9559	NNCUGGGUGNN	(((.....)))	3.78E-06	0.0131605
Ptm9654	YUUNGGCAAG	(((....)))	3.90E-06	3.90E-06
Ptm9676	YUGAGCAG	((....))	3.93E-06	3.93E-06
Ptm9692	UUCGCCNGRR	(((....)))	3.95E-06	3.95E-06
Ptm9710	NCGGACCNGN	((......))	3.97E-06	0.000174403
Ptm9751	NGGUCGCNNNRYYN	((((......))))	4.01E-06	0.000143025
Ptm9755	ANCGNGNNYGNU	((((....))))	4.01E-06	0.00015403
Ptm9862	NGAGGACYUYN	((((...))))	4.18E-06	0.000154959
Ptm9912	NNRNGGNNUCCNUNN	((((((...))))))	4.28E-06	0.194017
Ptm9966	NNGGAUAANNYYNN	((((......))))	4.36E-06	0.0037179
Ptm10028	RGUGNNNNCCACU	((((.....))))	4.46E-06	4.46E-06

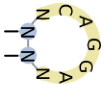
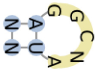
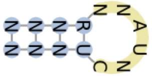
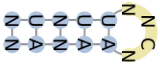
Ptm10064	AUNCGNNNNYGNRU	(((((....)))))	4.51E-06	0.00132031
Ptm10308	NGGACAUNNNNUYYN	((((.......))))	4.83E-06	0.000113459
Ptm10419	RUYNUCCGAU	(((....)))	5.03E-06	5.03E-06
Ptm10430	NCAAUGANUGN	(((.....)))	5.05E-06	0.000183325
Ptm10436	GYGAGACGC	(((...)))	5.06E-06	5.06E-06
Ptm10471	GNUGCANNNNNUGYRNY	((((((.....))))))	5.10E-06	1.59E-08
Ptm10472	UCGGCCNNNNYGR	(((.......)))	5.10E-06	5.10E-06
Ptm10555	CGNCCGNGNYG	((((...))))	5.26E-06	0.00030061
Ptm10574	YGGCCACCG	(((...)))	5.30E-06	5.30E-06
Ptm10592	NNUNUGGUCANN	(((......)))	5.32E-06	0.012096
Ptm10871	NNNNGRRNNNNANNUUCNNNN	(((((((.......)))))))	5.72E-06	0.0547071
Ptm10877	GUYNNNNGUGGAC	(((.......)))	5.74E-06	5.74E-06
Ptm10978	YUNGNNGGCNAG	((((....))))	5.90E-06	0.000159964
Ptm11049	NCCGNCUNNYGGN	((((.....))))	6.02E-06	0.000263417
Ptm11124	GYYRYNNNGGUGGC	(((((....)))))	6.12E-06	6.12E-06
Ptm11132	NYGNNNUCCCCGN	(((.......)))	6.14E-06	0.000486648
Ptm11205	GGAUNNNNNNNNNNNNNRUYY	(((((((.......)))))))	6.26E-06	0.000331528
Ptm11529	CUGGCUNYRG	(((....)))	6.85E-06	6.85E-06
Ptm11628	GRRUNNNCNAUUC	((((.....))))	7.04E-06	7.04E-06
Ptm11636	NNNNNNNNNCCCGNNNNNN	((((((.......))))))	7.06E-06	0.00926905
Ptm11698	GRYYNNGGGUC	((((...))))	7.20E-06	7.20E-06
Ptm11997	NYYRNYNNNNNNAGNUGGN	((((((.......))))))	7.75E-06	0.00780217
Ptm12076	NGYNNUNNNNNCANNGCN	((((((......))))))	7.92E-06	0.39183
Ptm12102	NNNNGRNANNAUCNNNN	((((((.....))))))	7.97E-06	0.000565038
Ptm12122	NUGYNNNGCAGCAN	((((......))))	8.02E-06	0.000394556
Ptm12343	NNNGUCNGANRYNNN	(((((.....)))))	8.50E-06	4.11E-05
Ptm12496	GUUNNNNNGNNAAC	(((((....)))))	8.85E-06	6.72E-05
Ptm12558	NNNNGAGNNGNYUYNNNN	(((((((....)))))))	8.95E-06	0.00022996
Ptm12631	YNGNNNUGCCNG	(((......)))	9.10E-06	3.45E-08
Ptm12697	AGACGGNNU	(.......)	9.27E-06	9.27E-06
Ptm12831	NGCUNUUNNNRNRGYN	((((((....))))))	9.52E-06	0.00848764
Ptm12841	GYUYNNNNAGGAGC	((((......))))	9.56E-06	9.56E-06
Ptm12948	UUGUUGNNNNNNNYRRYRR	(((((((.....)))))))	9.80E-06	0.000306094
Ptm12969	GYRRGNNNCUUGC	(((((...)))))	9.84E-06	9.84E-06
Ptm13050	NYNGAUACNGN	((((...))))	1.00E-05	0.000196313
Ptm13101	CGUGCUNNNNNYRYG	((((.......))))	1.01E-05	1.01E-05
Ptm13118	NNRYNAACAGUNN	((((.....))))	1.02E-05	0.0156895
Ptm13506	YGUNNNNGUCACG	(((.......)))	1.12E-05	1.12E-05
Ptm13522	AACNUANNNNGUU	(((.......)))	1.12E-05	1.12E-05
Ptm13677	NNNGAACUNUYNNN	(((((....)))))	1.17E-05	0.188556
Ptm13701	NNNNUCNAGURNNNN	(((((.....)))))	1.17E-05	0.438242

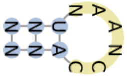
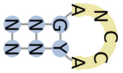
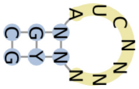
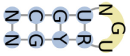
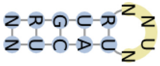
Ptm13749	GACAACC	(.....)	1.19E-05	1.19E-05
Ptm13832	NAACAUGUUN	(((....)))	1.21E-05	0.000352276
Ptm14266	NGGUCAANYYN	(((.....)))	1.31E-05	0.00053539
Ptm14286	NGGANUCNNUYYN	(((((...)))))	1.32E-05	0.0139244
Ptm14482	NUYRNCAUUGAN	((((....))))	1.38E-05	0.000442055
Ptm14520	GYNNAGAAGC	((......))	1.39E-05	1.39E-05
Ptm14585	NCNAGNCUNGN	((((...))))	1.41E-05	0.000249175
Ptm14648	NNNAUNAGGNNUNNN	((((.......))))	1.43E-05	0.0183699
Ptm14683	NGCACUANNNGYN	(((.......)))	1.44E-05	0.000342053
Ptm14703	NNGGUNNCNNNNNRYYNN	(((((((....)))))))	1.44E-05	0.0345405
Ptm14755	NNURNNNGUCNUANN	((((.......))))	1.46E-05	0.0203366
Ptm14907	CAUGACRUG	(((...)))	1.51E-05	1.51E-05
Ptm14967	NYYNRYNNNNNCGUNGGN	((((((......))))))	1.52E-05	0.00704939
Ptm15195	NGAUCCNRUYN	((((...))))	1.59E-05	0.000563353
Ptm15269	NNYGUNNNNNCNNACGNN	((((((......))))))	1.61E-05	0.046329
Ptm15402	NNGNNCAGCGCNN	(((.......)))	1.65E-05	0.022509
Ptm15491	NNNGYRNNNNNGGUGCNNN	((((((.......))))))	1.68E-05	0.034878
Ptm15495	GYRRNCUUUGC	((((...))))	1.68E-05	1.68E-05
Ptm15588	YGNUNGAANCG	((((...))))	1.70E-05	0.000693173
Ptm15706	AAGGCNYUU	(((...)))	1.73E-05	1.73E-05
Ptm15809	ACUUNANNNRRGU	((((.....))))	1.78E-05	1.78E-05
Ptm15823	NURRYNANGUUAN	(((((...)))))	1.78E-05	0.000617086
Ptm15869	GUCGCANNNNGRY	(((.......)))	1.79E-05	1.79E-05
Ptm15987	NUGGACNNNNUYYRN	(((((.....)))))	1.83E-05	4.20E-07
Ptm15998	NUNCCGNCAN	((......))	1.84E-05	0.00196215
Ptm16101	NNACUUCANNUNN	(((.......)))	1.87E-05	0.0719592
Ptm16121	NNNNGAUUACNNN	(((.......)))	1.87E-05	0.000141372
Ptm16238	UANGACNNNNYNUR	((((......))))	1.91E-05	0.000839546
Ptm16614	NNRRNNNUGGAUUNN	((((.......))))	2.06E-05	0.0189975
Ptm16983	NNGUANCGNURYNN	(((((....)))))	2.19E-05	0.0159233
Ptm17098	AGCAGNNUGYU	((((...))))	2.23E-05	2.23E-05
Ptm17183	GYYNYYNNNNNNGGNGGC	((((((......))))))	2.27E-05	0.000650286
Ptm17203	NNYRNNNCAACUGNN	((((.......))))	2.27E-05	0.0441853
Ptm17349	NNNNNNUCUGNUNNNNN	(((((.......)))))	2.33E-05	0.516376
Ptm17370	AUCUCGRU	((....))	2.33E-05	2.33E-05
Ptm17476	AGUACANRYU	(((....)))	2.38E-05	2.38E-05
Ptm17656	UGYACGGCA	(((...)))	2.45E-05	2.45E-05
Ptm17670	NNNAANCNAUUNNN	(((((....)))))	2.46E-05	0.159815
Ptm17727	NNNANAGCCNNUNNN	((((.......))))	2.48E-05	0.0304499
Ptm17771	NCUGGCANNRGN	(((......)))	2.50E-05	0.00110732
Ptm17941	NNUUNNNGCGNAANN	((((.......))))	2.58E-05	0.0302369

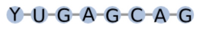
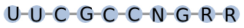
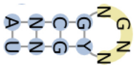
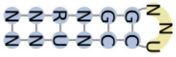
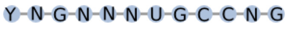
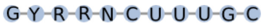
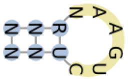
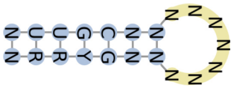
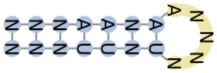
Ptm17945	NNNYYNGGUNGGNNN	(((((.....)))))	2.58E-05	0.0639509
Ptm17951	NNACANCAGUNN	((((....))))	2.58E-05	0.0522543
Ptm18023	GGYNUANGCC	(((....)))	2.61E-05	2.61E-05
Ptm18026	NCAGGCNNNNYUGN	((((......))))	2.61E-05	0.000754074
Ptm18467	AGGCNGNNNGYYU	((((.....))))	2.80E-05	2.80E-05
Ptm18494	GCCUNUNNNNNRGGY	((((.......))))	2.81E-05	2.81E-05
Ptm18807	UNNAAGNNNNNNNYUUNNR	((((((.......))))))	2.95E-05	0.0291851
Ptm18974	GAAUUANNNNUUY	(((.......)))	3.03E-05	3.03E-05
Ptm19174	GUYRNNNNNAUGAC	((((......))))	3.13E-05	3.13E-05
Ptm19396	NNNNNRYNNNNANGGUNNNNN	(((((((.......)))))))	3.23E-05	0.00321787
Ptm19412	NAUGGCANRUN	(((.....)))	3.24E-05	0.00105029
Ptm19445	GURNNNNCANUAC	((((.....))))	3.25E-05	0.000892911
Ptm19544	NCNGAGCNNNNYNGN	((((.......))))	3.30E-05	0.0406554
Ptm19642	NUYNNGCGGGAN	(((......)))	3.36E-05	0.00135465
Ptm19982	NNYUNYNNNGNGNAGNN	(((((((...)))))))	3.51E-05	0.0480054
Ptm19983	NUAGUUGNNURN	(((......)))	3.52E-05	0.00113791
Ptm20057	URNCUACUA	((.....))	3.56E-05	3.56E-05
Ptm20588	NUGGUUCYYRN	((((...))))	3.84E-05	5.93E-07
Ptm20807	NUYNAGGUGAN	(((.....)))	3.94E-05	0.00150963
Ptm20879	NNRNAAGUCUNN	(((......)))	3.98E-05	0.0332516
Ptm20955	NNYGNAUCCGNN	((((....))))	4.03E-05	0.0395853
Ptm21104	GGNNNUAUUCC	((.......))	4.12E-05	4.12E-05
Ptm21126	ACAGGCNNGU	((......))	4.13E-05	4.13E-05
Ptm21293	NNNCNCAGCNNN	(((......)))	4.23E-05	4.89E-07
Ptm21328	NUUGCNNNNNNNNNNNGYRRN	(((((((.......)))))))	4.24E-05	7.18E-05
Ptm21454	YYRUNNNANAUGG	(((((...)))))	4.33E-05	8.12E-07
Ptm21561	GRRNNNAAGUUC	(((......)))	4.39E-05	4.39E-05
Ptm21745	NNGGUGCGNNNYYNN	((((.......))))	4.49E-05	0.000836763
Ptm22145	NNNAANAANNNNUNUUNNN	(((((((.....)))))))	4.74E-05	0.0610512

Supplementary Table 9-2. Predicted motifs that are supported by analysis of previously published microarray/RNA-seq data or by microarray analysis of chemical perturbations reported in this study – The structure of a motif is shown only when the *p*-value of conservation is better with the structure compared to the sequence alone. Otherwise, the motif is shown as a linear sequence.

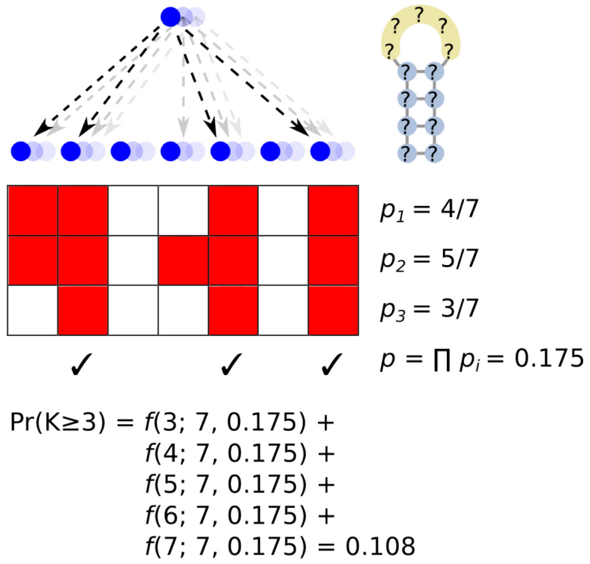
Motif	Sequence/structure	Conservation *p*-value (sequence and structure)	Conservation *p*-value (sequence only)	Notes
Ptm1	C-A-U-A-G-A-N	2.12×10^{-29}	2.12×10^{-29}	• Known binding site for cycling sequence-binding proteins • Up-regulated in stationary-phase PF • Co-regulated across different experiments
Ptm13	U-N-A-U-G-G-A	1.43×10^{-19}	1.43×10^{-19}	• Down-regulated in stationary-phase PF • Co-regulated across different experiments
Ptm15	U-Y-G-A-N-G-A	1.64×10^{-18}	1.64×10^{-18}	• Up-regulated in stumpy BF • Down-regulated in EtBr-treated PF • Up-regulated in HCl-treated PF • Co-regulated across different experiments
Ptm17	A-U-G-U-A-N	3.65×10^{-18}	3.65×10^{-18}	• Down-regulated in stumpy BF • Up-regulated in stationary-phase PF • Up-regulated in PARN1-overexpressing PF • Up-regulated in EtBr-treated PF • Down-regulated in HCl-treated PF • Down-regulated in verapamil-treated PF • Co-regulated across different experiments
Ptm23	U-Y-G-C-N-G-A	1.06×10^{-17}	1.06×10^{-17}	• Up-regulated in slender BF • Down-regulated in stationary-phase PF • Down-regulated in PARN1-overexpressing PF • Down-regulated in EtBr-treated PF • Up-regulated in HCl-treated PF • Co-regulated across different experiments
Ptm25	U-Y-C-G-N-G-A	3.95×10^{-17}	3.95×10^{-17}	• Up-regulated in HCl-treated PF • Co-regulated across different experiments

Ptm39	G-C-N-C-C-N-N-N-N-G-Y	5.02×10^{-16}	5.02×10^{-16}	• Down-regulated in stationary-phase PF • Down-regulated in PARN1-overexpressing PF • Down-regulated in DHH1-overexpressing PF • Down-regulated in EtBr-treated PF • Up-regulated in DMSO-treated PF • Up-regulated in HCl-treated PF • Down-regulated in hygromycin-treated PF • Co-regulated across different experiments
Ptm80		9.42×10^{-14}	9.54×10^{-07}	• Down-regulated in PARN1-overexpressing PF • Down-regulated in EtBr-treated PF • Up-regulated in HCl-treated PF • Co-regulated across different experiments
Ptm127	C-N-A-A-G-G-N-N-G	1.24×10^{-12}	1.24×10^{-12}	• Down-regulated in DHH1-overexpressing PF • Down-regulated in mutant DHH1-expressing PF • Up-regulated in NaOH-treated PF • Co-regulated across different experiments
Ptm176	U-G-G-C-C-N-N-N-Y-Y-R	4.81×10^{-12}	4.81×10^{-12}	• Co-regulated across different experiments
Ptm234		2.12×10^{-11}	6.33×10^{-09}	• Down-regulated in *in vitro*-cultured BF • Co-regulated across different experiments
Ptm360	G-G-U-N-N-A-N-U-A-C-C	1.07×10^{-10}	1.07×10^{-10}	• Down-regulated in stationary-phase PF • Co-regulated across different experiments
Ptm509		3.53×10^{-10}	0.002543	• Co-regulated across different experiments
Ptm580	A-C-U-G-G-A-R-G-U	5.46×10^{-10}	5.46×10^{-10}	• Down-regulated in stationary-phase PF
Ptm1374		9.19×10^{-09}	0.000425	• Up-regulated in DHH1-overexpressing PF

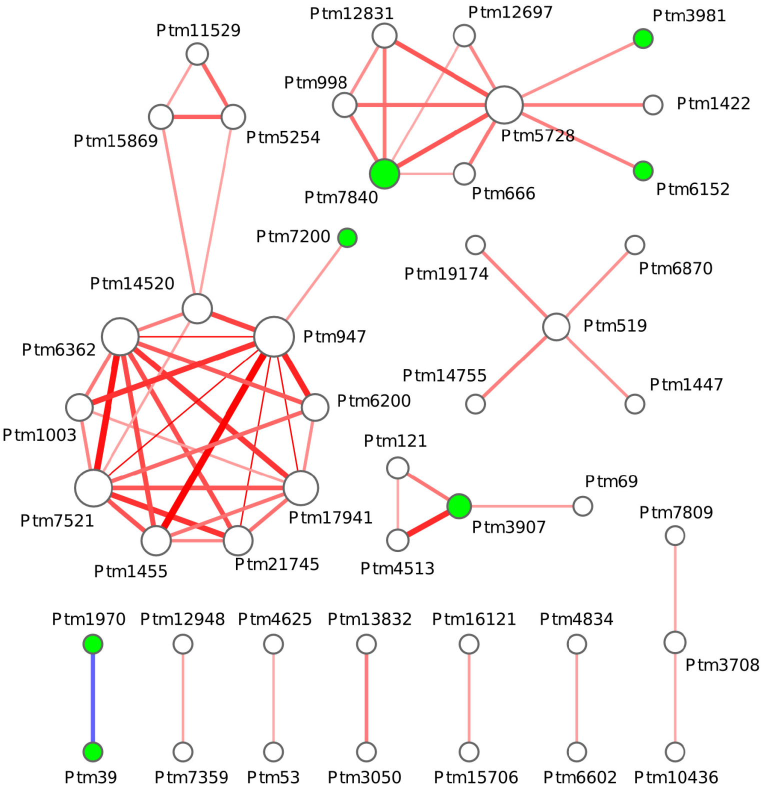
Ptm1970	A-U-U-U-A-U-U	2.86×10^{-08}	2.86×10^{-08}	• Down-regulated in stumpy and slender BF • Up-regulated in stationary-phase PF • Up-regulated in PARN1-overexpressing PF • Up-regulated in DHH1-overexpressing PF • Up-regulated in EtBr-treated PF • Down-regulated in DMSO-treated PF • Down-regulated in HCl-treated PF • Down-regulated in verapamil-treated PF • Co-regulated across different experiments
Ptm2405		5.29×10^{-08}	2.05×10^{-05}	• Down-regulated in slender vs. stumpy BF
Ptm2447		5.58×10^{-08}	0.194717	• Down-regulated in hygromycin-treated PF
Ptm3907	Y U U-Y-R-N-N-A-U-G-A-A-G	2.46×10^{-07}	2.46×10^{-07}	• Down-regulated in stationary-phase PF • Up-regulated in mutant DHH1-expressing PF • Down-regulated in DMSO-treated PF • Co-regulated across different experiments
Ptm3981		2.61×10^{-07}	8.76×10^{-06}	• Co-regulated across different experiments
Ptm4905	Y-U-G-Y-R-N-N-C-U-G-C-A-G	4.97×10^{-07}	4.97×10^{-07}	• Up-regulated in verapamil-treated PF
Ptm5069	C-A-G-G-G-N-N-N-Y-U-G	5.53×10^{-07}	5.53×10^{-07}	• Co-regulated across different experiments
Ptm6152		9.82×10^{-07}	2.96×10^{-05}	• Up-regulated in verapamil-treated PF
Ptm7127	A-G-U-C-N-G-N-N-N-N-R-Y-U	1.53×10^{-06}	1.53×10^{-06}	• Co-regulated across different experiments
Ptm7200	G-N-C-U-C-N-A-C	1.57×10^{-06}	1.57×10^{-06}	• Up-regulated in HCl-treated PF
Ptm7840		2.06×10^{-06}	6.20×10^{-05}	• Down-regulated in slender vs. stumpy BF • Co-regulated across different experiments

Ptm9676		3.93×10^{-06}	3.93×10^{-06}	• Co-regulated across different experiments
Ptm9692		3.95×10^{-06}	3.95×10^{-06}	• Co-regulated across different experiments
Ptm9755		4.01×10^{-06}	0.000154	• Down-regulated in Alba3/4 RNAi in PF • Co-regulated across different experiments
Ptm9912		4.28×10^{-06}	0.194017	• Down-regulated in DHH1-overexpressing PF
Ptm12631		9.10×10^{-06}	3.45×10^{-08}	• Down-regulated in stationary-phase PF
Ptm15495		1.68×10^{-05}	1.68×10^{-05}	• Up-regulated in slender BF • Co-regulated across different experiments
Ptm20879		3.98×10^{-05}	0.033252	• Up-regulated in Alba3/4 RNAi in PF
Ptm21328		4.24×10^{-05}	7.18×10^{-05}	• Up-regulated in NaOH-treated PF
Ptm22145		4.74×10^{-05}	0.061051	• Up-regulated in stationary-phase PF • Co-regulated across different experiments

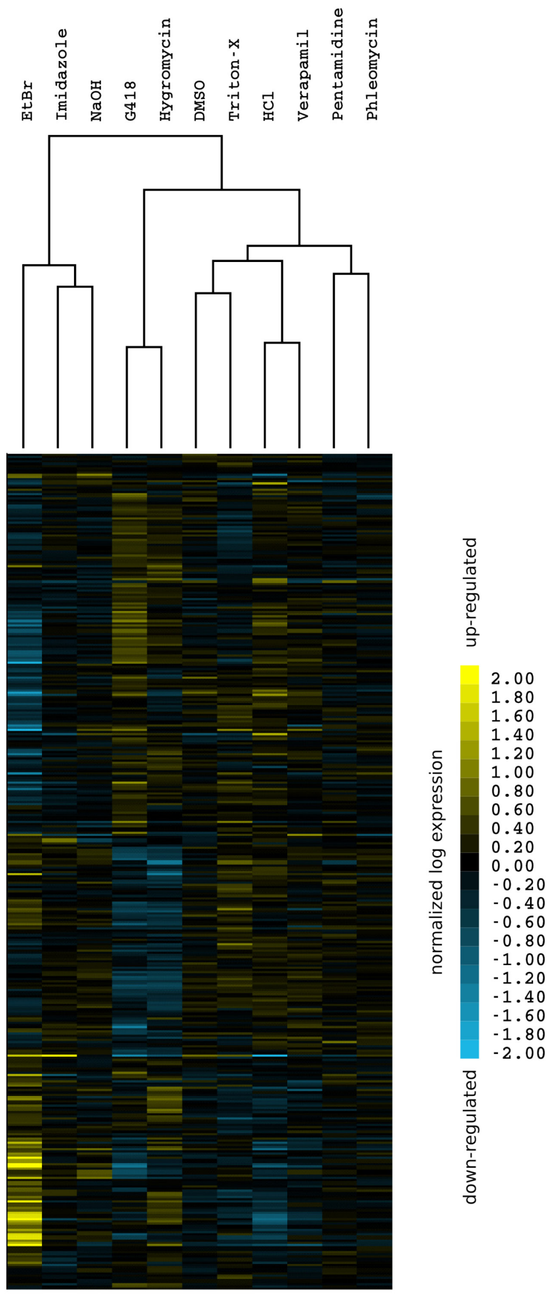
9.6 Supplementary Figures



Supplementary Figure 9-1. Finding motifs with high network-level conservation – COSMOS (Conserved Structural Motif Search tool) uses a binomial distribution-based statistical framework in order to identify motifs that are highly conserved at network level, i.e. motifs that correspond to the binding sites of conserved trans-regulatory factors with conserved regulatory networks (top of the figure). Given a particular motif with sequence/structure information, COSMOS estimates the probability of occurrence of the motif in a regulatory region in each organism (*p_1*, *p_2* and *p_3*), and then uses them to predict the probability (*p*) of observing a "keeper" (shown by the check marks), i.e. an ortholog group in which all genes contain at least one instance of that motif in their regulatory region. Then, using the binomial distribution, COSMOS estimates the probability of the observed number of motifs under the null hypothesis of random motif distribution. Small probability values correspond to motifs that are highly conserved at network level.



Supplementary Figure 9-2. Regulatory modules in *T. brucei*. We identified 60 positive interactions and one negative interaction among 46 predicted conserved motifs in *T. brucei*. For each pair of motifs, functional interaction was identified based on the extent of overlap/exclusiveness of their carrier transcripts. The overlap of target genes was measured by hypergeometric distribution at Bonferroni-corrected *p*-value cutoff of 0.025. In this figure, red edges indicate positive interactions (significant overlap between the target sets of two motifs) and the blue edge represents negative interaction (an exclusive pair of motif). The overlap/exclusion *p*-value is shown by the thickness of the edges, with more significant interactions represented by thicker edges. The node size represents the number of functional interactions (degree) for each motif. Green nodes stand for motifs that were validated in this study based on previous expression data or based on microarray analysis of chemical perturbations.



Supplementary Figure 9-3. Expression profiling of chemical perturbations in *T. brucei* – Each row represents one gene, and each column one experiment. Columns are normalized to have an average of 0 and standard deviation of 1. Yellow and blue represent up- and down-regulation, respectively. For visualization purposes, genes with missing values are not shown here. Missing values were because of low signal-to-noise ratio and, thus, unreliable measurements in all probes corresponding to a gene.

10 Concluding remarks and future directions

As I write this last chapter of my thesis, 1738 complete genome sequences are available on NCBI [342], 468 genomes are in the assembly stage, 707 genomes are being sequenced, and UniProtKB [343] holds over 17,000,000 protein entries, of which more than 5,000,000 are simply named “Uncharacterized protein”. How are we going to deal with this massive amount of data? Computational methods to analyze and annotate nucleotide and protein sequences are now far behind the rapidly growing mass of sequence data, and methods to extract knowledge are even scarcer. The focus of my thesis has been on new methods for functional annotation of nucleotide and protein sequences. We have developed several methods that can predict the functions of genes based on their coding and translated sequences, as well as based on the sequences of their regulatory regions. These methods are aimed to be “homology-independent”, meaning that they can functionally annotate non-conserved genes. We have applied these methods to annotate uncharacterized proteins of *Trypanosoma brucei*, a parasite that causes the devastating human African trypanosomiasis and takes tens of thousands of victims every year: just in the year 2008, the first year of my PhD studies, 48000 people died of African trypanosomiasis [344].

Availability of the genome sequences of trypanosomatids [25-29], including *T. brucei*, has transformed the research on the rather neglected diseases that they create. At this moment, there are more than 600 scholarly articles and books that have cited the genome sequence of *T. brucei*, clearly showing the extent to which researchers are using these data to explore the biology of this parasite. Recently, high-throughput techniques for genome-wide identification of essential genes in *T. brucei* have been in the spotlight [345], which may prove the next major leap towards identification of drug targets and development of new therapeutics. Yet, we do not know the function of the majority of trypanosomatid proteins, and until we do, our options for drug targets are limited.

This thesis has described a summary of our efforts towards functional characterization of *T. brucei* genes and proteins, and towards understanding how they are regulated within the cell. The scarcity of functional genomics data made us explore different possibilities, such as using codon usage for function prediction, identification of function-specific regulatory elements without using expression data, and finding conserved regulatory programs when regulatory regions cannot be aligned. These developments have provided the material for creating a computational pipeline that enables us to functionally annotate genomes using the concept of homology-independent sequence-based annotation. The integration of different methods and conceptual advances that are presented in this thesis will be the next natural step towards building this computational pipeline. For example, Bayesian networks can provide suitable frameworks for integration of different computational methods. They have been previously used for integration of different datasets in order to predict protein-protein interactions [81], and we have also successfully used them to integrate codon usage and function-specific *cis*-regulatory elements in order to predict gene function [30]. As we proposed in Chapter 7, this framework should ideally create an automated pipeline for annotation of any newly sequenced genome.

In addition to development of new computational tools, there is an obvious need for more comprehensive functional genomics data of *T. brucei*, i.e. data that uncover dynamic aspects of gene function in the cell, such as transcriptome and proteome profiles across different conditions as well as protein and genetic interaction networks. In Chapter 8 of this thesis, we analyzed the available gene expression datasets of *T. brucei* and *L. infantum*, and demonstrated the conserved co-expression of functionally linked genes in trypanosomatids. However, the co-expression network that we obtained was far from complete, and could only partially uncover functional relationships among genes. The low coverage of this network primarily stemmed from the very small number of expression datasets that were available at the time: we used 17 expression datasets, of which eight seemed to provide little useful information about functional linkages. Identification of co-expressed genes from such a small number of datasets is a challenging task, and results are often noisy and unreliable. Similar studies of co-expression networks in organisms such as mouse, yeast and human [295, 346, 347] have used hundreds of genome-wide expression profiles, and researchers have even gone farther and have combined the data across multiple species to obtain a global view of conserved genetic modules [283].

Currently, the number of available non-redundant expression datasets of *T. brucei* has grown to about 25, a moderate growth compared to the 17 datasets that were available in 2010. Furthermore, except for a very few, almost all these datasets have examined the developmental gene regulation in *T. brucei*, ignoring gene regulation within a life stage in response to environmental and internal stimuli. Our microarray analysis of chemical perturbations, presented in Chapter 9, clearly showed that *T. brucei* transcriptome remodels extensively upon perturbation of pathways and biological processes. Thus, a comprehensive analysis of different perturbations may prove very useful in identifying genetic modules and, subsequently, functionally related genes. For this purpose, one can use a wide spectrum of chemicals to perturb different biological processes, each perturbation triggering a specific transcriptome response, which will eventually reveal sets of co-regulated genes. This approach has been successfully used in the past to identify sets of co-regulated genes in yeast [347].

A complementary approach is to identify epistatic interactions among genes. Genes that are in the same pathway, in the same protein complex, or in parallel pathways often have genetic interactions, meaning that the function or effect of one gene depends upon or is modified by the function of another gene. Genetic interactions can be used to identify functional linkages among genes, leading to identification of pathways and biological modules [348]. Furthermore, changes that occur in genetic interactions, as a result of environmental modifications, are highly informative about the functions of genes [349]. Genetic interactions are conventionally reconstructed by analysis of a large number of double knockout/knockdown strains, with the aim of identifying pairs of genes whose double inhibition results in a phenotype that is significantly stronger or weaker than would be expected based on combined phenotypes of single gene inhibitions. In *T. brucei*, we have the unique opportunity of using the RNAi machinery for efficient inhibition of genes. Double gene inhibition can be easily achieved in this organism by using chimerical RNAi constructs that carry fragments from two different genes [350], making it possible to conduct focused genetic interaction studies in *T. brucei*. Also, combining high-throughput sequencing with double RNAi libraries may provide us with necessary tools to construct a genome-wide genetic interaction map of *T. brucei*, which will lead us to a global picture of gene functions in this organism. A recent study has used high-throughput sequencing for massively parallel phenotyping of single RNAi libraries [345]. With minor modifications, the same approach can be adapted for massively parallel phenotyping of highly complex double RNAi libraries, especially that high-throughput sequencing technologies allow paired-end sequencing of DNA fragments.

Lastly, I believe that the key to understanding the behavior of trypanosomatids is in understanding the mechanisms they use to regulate their genes and biological processes. *T. brucei* genome contains more than 220 RNA-binding proteins (RBPs), suggesting the presence of a complex post-transcriptional network in this parasite. Our computational analysis of conserved regulatory programs, as presented in Chapter 9, revealed potential *cis*- and *trans*-acting regulatory factors in this organism. However, this was just a first step towards characterization of the gene regulatory network of *T. brucei*, and should be followed by extensive experimental characterization of the many uncharacterized RNA-binding proteins. Again, RNAi provides a unique opportunity for deciphering the regulatory code in *T. brucei*. Previously, it has been shown that RNAi-mediated inhibition of *trans*-acting post-transcriptional regulatory proteins has a measurable effect on the level of their target mRNAs [351]. Therefore, comprehensive RNAi-mediated inhibition of known and predicted RBPs followed by transcriptome profiling (e.g. using microarrays) can be used to identify the targets of RBPs, as well as to understand the stabilizing or destabilizing effect of the RBPs on their target transcripts. This is in contrast to alternative approaches such as RIP-chip and CLIP-seq [352, 353], which identify only the binding targets and not the effect of the RBPs on their targets. Such a comprehensive analysis of RBPs will also contribute significantly to prediction of gene functions in *T. brucei*, as we expect functionally linked genes to be regulated by a common set of regulatory elements [3].

11 Contribution to knowledge

This thesis presents a battery of novel computational and experimental approaches in order to discover the functions of uncharacterized genes, as well as to identify the regulatory mechanisms that regulate genes and pathways. We have applied these methods to characterize the functions and regulatory codes of genes in the parasitic protozoan *Trypanosoma brucei*. However, the methods that are presented here are not limited to this organism, and can be used to functionally characterize coding and non-coding sequences of all organisms. Accordingly, the contribution of this thesis to knowledge can be summarized as the following:

1. Discovering a novel role for codon usage in dynamic regulation of protein expression;
2. Introducing novel computational methods for sequence-based homology-independent prediction of biological processes and pathways as well as protein molecular functions;
3. Discovering the biological processes and pathways of more than 100 previously uncharacterized *T. brucei* genes;
4. Discovering the molecular functions of more than 2600 previously uncharacterized *T. brucei* proteins;
5. Introducing novel computational methods for global characterization of conserved post-transcriptional regulatory programs;
6. Discovering and characterizing 35 novel *cis*-regulatory elements and four novel *trans*-regulatory elements in *T. brucei*.

12 References

1. Najafabadi HS, Salavati R: **Sequence-based prediction of protein-protein interactions by means of codon usage.** *Genome biology* 2008, **9:**R87.

2. Najafabadi HS, Goodarzi H, Salavati R: **Universal function-specificity of codon usage.** *Nucleic Acids Res* 2009, **37:**7014-7023.

3. Mao Y, Najafabadi HS, Salavati R: **Genome-wide computational identification of functional RNA elements in Trypanosoma brucei.** *BMC genomics* 2009, **10:**355.

4. Salavati R, Najafabadi HS: **Sequence-based functional annotation: what if most of the genes are unique to a genome?** *Trends Parasitol* 2010, **26:**225-229.

5. Shateri Najafabadi H, Salavati R: **Functional genome annotation by combined analysis across microarray studies of Trypanosoma brucei.** *PLoS neglected tropical diseases* 2010, **4**.

6. Vickerman K: **The evolutionary expansion of the trypanosomatid flagellates.** *Int J Parasitol* 1994, **24:**1317-1331.

7. **WHO - Global Burden of Disease Estimates** [http://www.who.int/healthinfo/bodestimates/en/index.html]

8. Stuart K, Brun R, Croft S, Fairlamb A, Gurtler RE, McKerrow J, Reed S, Tarleton R: **Kinetoplastids: related protozoan pathogens, different diseases.** *J Clin Invest* 2008, **118:**1301-1310.

9. Teixeira AR, Nitz N, Guimaro MC, Gomes C, Santos-Buch CA: **Chagas disease.** *Postgrad Med J* 2006, **82:**788-798.

10. Delespaux V, de Koning HP: **Drugs and drug resistance in African trypanosomiasis.** *Drug Resist Updat* 2007, **10:**30-50.

11. Croft SL, Barrett MP, Urbina JA: **Chemotherapy of trypanosomiases and leishmaniasis.** *Trends Parasitol* 2005, **21:**508-512.

12. den Boer M, Davidson RN: **Treatment options for visceral leishmaniasis.** *Expert Rev Anti Infect Ther* 2006, **4:**187-197.

13. Olliaro PL, Guerin PJ, Gerstl S, Haaskjold AA, Rottingen JA, Sundar S: **Treatment options for visceral leishmaniasis: a systematic review of clinical studies done in India, 1980-2004.** *Lancet Infect Dis* 2005, **5:**763-774.

14. Buckner FS, Wilson AJ, White TC, Van Voorhis WC: **Induction of resistance to azole drugs in Trypanosoma cruzi.** *Antimicrob Agents Chemother* 1998, **42:**3245-3250.

15. Wilkinson SR, Taylor MC, Horn D, Kelly JM, Cheeseman I: **A mechanism for cross-resistance to nifurtimox and benznidazole in trypanosomes.** *Proc Natl Acad Sci U S A* 2008, **105:**5022-5027.

16. Cardo LJ: **Leishmania: risk to the blood supply.** *Transfusion* 2006, **46:**1641-1645.

17. Ozcan D, Seckin D, Allahverdiyev AM, Weina PJ, Aydin H, Ozcay F, Haberal M: **Liver transplant recipient with concomitant cutaneous and visceral leishmaniasis.** *Pediatr Transplant* 2007, **11:**228-232.

18. Riera C, Fisa R, Lopez-Chejade P, Serra T, Girona E, Jimenez M, Muncunill J, Sedeno M, Mascaro M, Udina M, et al: **Asymptomatic infection by Leishmania infantum in blood donors from the Balearic Islands (Spain).** *Transfusion* 2008, **48:**1383-1389.

19. Bern C, Montgomery SP, Katz L, Caglioti S, Stramer SL: **Chagas disease and the US blood supply.** *Curr Opin Infect Dis* 2008, **21:**476-482.

20. Kun H, Moore A, Mascola L, Steurer F, Lawrence G, Kubak B, Radhakrishna S, Leiby D, Herron R, Mone T, et al: **Transmission of Trypanosoma cruzi by heart transplantation.** *Clin Infect Dis* 2009, **48:**1534-1540.

21. Keynan Y, Larios OE, Wiseman MC, Plourde M, Ouellette M, Rubinstein E: **Use of oral miltefosine for cutaneous leishmaniasis in Canadian soldiers returning from Afghanistan.** *Can J Infect Dis Med Microbiol* 2008, **19:**394-396.

22. Faulde M, Schrader J, Heyl G, Amirih M: **Differences in transmission seasons as an epidemiological tool for characterization of anthroponotic and zoonotic cutaneous leishmaniasis in northern Afghanistan.** *Acta Trop* 2008, **105:**131-138.

23. Leslie T, Saleheen S, Sami M, Mayan I, Mahboob N, Fiekert K, Lenglet A, Ord R, Reithinger R: **Visceral leishmaniasis in Afghanistan.** *CMAJ* 2006, **175:**245.

24. Aronson NE, Sanders JW, Moran KA: **In harm's way: infections in deployed American military forces.** *Clin Infect Dis* 2006, **43:**1045-1051.

25. El-Sayed NM, Myler PJ, Bartholomeu DC, Nilsson D, Aggarwal G, Tran AN, Ghedin E, Worthey EA, Delcher AL, Blandin G, et al: **The genome sequence of Trypanosoma cruzi, etiologic agent of Chagas disease.** *Science* 2005, **309:**409-415.

26. El-Sayed NM, Myler PJ, Blandin G, Berriman M, Crabtree J, Aggarwal G, Caler E, Renauld H, Worthey EA, Hertz-Fowler C, et al: **Comparative genomics of trypanosomatid parasitic protozoa.** *Science* 2005, **309:**404-409.

27. Peacock CS, Seeger K, Harris D, Murphy L, Ruiz JC, Quail MA, Peters N, Adlem E, Tivey A, Aslett M, et al: **Comparative genomic analysis of three Leishmania species that cause diverse human disease.** *Nat Genet* 2007, **39:**839-847.

28. Berriman M, Ghedin E, Hertz-Fowler C, Blandin G, Renauld H, Bartholomeu DC, Lennard NJ, Caler E, Hamlin NE, Haas B, et al: **The genome of the African trypanosome Trypanosoma brucei.** *Science* 2005, **309:**416-422.

29. Ivens AC, Peacock CS, Worthey EA, Murphy L, Aggarwal G, Berriman M, Sisk E, Rajandream MA, Adlem E, Aert R, et al: **The genome of the kinetoplastid parasite, Leishmania major.** *Science* 2005, **309:**436-442.

30. Salavati R, Najafabadi HS: **Sequence-based functional annotation: what if most of the genes are unique to a genome?** *Trends Parasitol* 2010, **In Press**.

31. Clayton C, Shapira M: **Post-transcriptional regulation of gene expression in trypanosomes and leishmanias.** *Mol Biochem Parasitol* 2007, **156:**93-101.

32. Haile S, Papadopoulou B: **Developmental regulation of gene expression in trypanosomatid parasitic protozoa.** *Curr Opin Microbiol* 2007, **10:**569-577.

33. Kramer S, Queiroz R, Ellis L, Hoheisel JD, Clayton C, Carrington M: **The RNA helicase DHH1 is central to the correct expression of many developmentally regulated mRNAs in trypanosomes.** *Journal of cell science* 2010, **123:**699-711.

34. Kabani S, Fenn K, Ross A, Ivens A, Smith TK, Ghazal P, Matthews K: **Genome-wide expression profiling of in vivo-derived bloodstream parasite stages and dynamic analysis of mRNA alterations during synchronous differentiation in Trypanosoma brucei.** *BMC genomics* 2009, **10:**427.

35. Jensen BC, Sivam D, Kifer CT, Myler PJ, Parsons M: **Widespread variation in transcript abundance within and across developmental stages of Trypanosoma brucei.** *BMC genomics* 2009, **10:**482.

36. Veitch NJ, Johnson PC, Trivedi U, Terry S, Wildridge D, MacLeod A: **Digital gene expression analysis of two life cycle stages of the human-infective parasite, Trypanosoma brucei gambiense reveals differentially expressed clusters of co-regulated genes.** *BMC genomics* 2010, **11:**124.

37. Utter CJ, Garcia SA, Milone J, Bellofatto V: **Poly(A)-specific Ribonuclease (PARN-1) function in stage-specific mRNA turnover in Trypanosoma brucei.** *Eukaryotic cell* 2011.

38. Nilsson D, Gunasekera K, Mani J, Osteras M, Farinelli L, Baerlocher L, Roditi I, Ochsenreiter T: **Spliced leader trapping reveals widespread alternative splicing patterns in the highly dynamic transcriptome of Trypanosoma brucei.** *PLoS pathogens* 2010, **6**.

39. Haanstra JR, Kerkhoven EJ, van Tuijl A, Blits M, Wurst M, van Nuland R, Albert MA, Michels PA, Bouwman J, Clayton C, et al: **A domino effect in drug action: from metabolic assault towards parasite differentiation.** *Molecular microbiology* 2011, **79:**94-108.

40. Pays E, Vanhamme L, Perez-Morga D: **Antigenic variation in Trypanosoma brucei: facts, challenges and mysteries.** *Curr Opin Microbiol* 2004, **7:**369-374.

41. Andrade LO, Andrews NW: **The Trypanosoma cruzi-host-cell interplay: location, invasion, retention.** *Nat Rev Microbiol* 2005, **3:**819-823.

42. Rodriguez A, Martinez I, Chung A, Berlot CH, Andrews NW: **cAMP regulates Ca2+-dependent exocytosis of lysosomes and lysosome-mediated cell invasion by trypanosomes.** *J Biol Chem* 1999, **274:**16754-16759.

43. McConville MJ, de Souza D, Saunders E, Likic VA, Naderer T: **Living in a phagolysosome; metabolism of Leishmania amastigotes.** *Trends Parasitol* 2007, **23:**368-375.

44. Herwaldt BL: **Leishmaniasis.** *Lancet* 1999, **354:**1191-1199.

45. Tripathi P, Singh V, Naik S: **Immune response to leishmania: paradox rather than paradigm.** *FEMS Immunol Med Microbiol* 2007, **51:**229-242.

46. Kissinger JC: **A tale of three genomes: the kinetoplastids have arrived.** *Trends Parasitol* 2006, **22:**240-243.

47. Benz C, Nilsson D, Andersson B, Clayton C, Guilbride DL: **Messenger RNA processing sites in Trypanosoma brucei.** *Mol Biochem Parasitol* 2005, **143:**125-134.

48. Hotz HR, Hartmann C, Huober K, Hug M, Clayton C: **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* 1997, **25:**3017-3026.

49. Irmer H, Clayton C: **Degradation of the unstable EP1 mRNA in Trypanosoma brucei involves initial destruction of the 3'-untranslated region.** *Nucleic Acids Res* 2001, **29:**4707-4715.

50. Schurch N, Furger A, Kurath U, Roditi I: **Contributions of the procyclin 3' untranslated region and coding region to the regulation of expression in bloodstream forms of Trypanosoma brucei.** *Mol Biochem Parasitol* 1997, **89:**109-121.

51. Furger A, Schurch N, Kurath U, Roditi I: **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* 1997, **17:**4372-4380.

52. Hehl A, Vassella E, Braun R, Roditi I: **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* 1994, **91:**370-374.

53. Vassella E, Probst M, Schneider A, Studer E, Renggli CK, Roditi I: **Expression of a major surface protein of Trypanosoma brucei insect forms is**

controlled by the activity of mitochondrial enzymes. *Mol Biol Cell* 2004, **15:**3986-3993.

54. Vassella E, Den Abbeele JV, Butikofer P, Renggli CK, Furger A, Brun R, Roditi I: **A major surface glycoprotein of trypanosoma brucei is expressed transiently during development and can be regulated post-transcriptionally by glycerol or hypoxia.** *Genes Dev* 2000, **14:**615-626.

55. Quijada L, Guerra-Giraldez C, Drozdz M, Hartmann C, Irmer H, Ben-Dov C, Cristodero M, Ding M, Clayton C: **Expression of the human RNA-binding protein HuR in Trypanosoma brucei increases the abundance of mRNAs containing AU-rich regulatory elements.** *Nucleic Acids Res* 2002, **30:**4414-4424.

56. Webb H, Burns R, Ellis L, Kimblin N, Carrington M: **Developmentally regulated instability of the GPI-PLC mRNA is dependent on a short-lived protein factor.** *Nucleic Acids Res* 2005, **33:**1503-1512.

57. Mishra KK, Holzer TR, Moore LL, LeBowitz JH: **A negative regulatory element controls mRNA abundance of the Leishmania mexicana Paraflagellar rod gene PFR2.** *Eukaryot Cell* 2003, **2:**1009-1017.

58. Purdy JE, Donelson JE, Wilson ME: **Regulation of genes encoding the major surface protease of Leishmania chagasi via mRNA stability.** *Mol Biochem Parasitol* 2005, **142:**88-97.

59. Teixeira SM, Kirchhoff LV, Donelson JE: **Post-transcriptional elements regulating expression of mRNAs from the amastin/tuzin gene cluster of Trypanosoma cruzi.** *J Biol Chem* 1995, **270:**22586-22594.

60. Coughlin BC, Teixeira SM, Kirchhoff LV, Donelson JE: **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* 2000, **275:**12051-12060.

61. Mayho M, Fenn K, Craddy P, Crosthwaite S, Matthews K: **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* 2006, **34:**5312-5324.

62. D'Orso I, Frasch AC: **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* 2002, **277:**50520-50528.

63. Di Noia JM, D'Orso I, Sanchez DO, Frasch AC: **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* 2000, **275:**10218-10227.

64. Bringaud F, Muller M, Cerqueira GC, Smith M, Rochette A, El-Sayed NM, Papadopoulou B, Ghedin E: **Members of a large retroposon family are determinants of post-transcriptional gene expression in Leishmania.** *PLoS Pathog* 2007, **3:**1291-1307.

65. McNicoll F, Muller M, Cloutier S, Boilard N, Rochette A, Dube M, Papadopoulou B: **Distinct 3'-untranslated region elements regulate stage-specific mRNA accumulation and translation in Leishmania.** *J Biol Chem* 2005, **280:**35238-35246.

66. Boucher N, Wu Y, Dumas C, Dube M, Sereno D, Breton M, Papadopoulou B: **A common mechanism of stage-regulated gene expression in Leishmania mediated by a conserved 3'-untranslated region element.** *J Biol Chem* 2002, **277:**19511-19520.

67. Engstler M, Boshart M: **Cold shock and regulation of surface protein trafficking convey sensitization to inducers of stage differentiation in Trypanosoma brucei.** *Genes Dev* 2004, **18:**2798-2811.

68. Colasante C, Robles A, Li CH, Schwede A, Benz C, Voncken F, Guilbride DL, Clayton C: **Regulated expression of glycosomal phosphoglycerate kinase in Trypanosoma brucei.** *Mol Biochem Parasitol* 2007, **151:**193-204.

69. Zilka A, Garlapati S, Dahan E, Yaolsky V, Shapira M: **Developmental regulation of heat shock protein 83 in Leishmania. 3' processing and mRNA stability control transcript abundance, and translation id directed by a determinant in the 3'-untranslated region.** *J Biol Chem* 2001, **276:**47922-47929.

70. Quijada L, Soto M, Alonso C, Requena JM: **Identification of a putative regulatory element in the 3'-untranslated region that controls expression of HSP70 in Leishmania infantum.** *Mol Biochem Parasitol* 2000, **110:**79-91.

71. Murray A, Fu C, Habibi G, McMaster WR: **Regions in the 3' untranslated region confer stage-specific expression to the Leishmania mexicana a600-4 gene.** *Mol Biochem Parasitol* 2007, **153:**125-132.

72. Sharan R, Ulitsky I, Shamir R: **Network-based prediction of protein function.** *Mol Syst Biol* 2007, **3:**88.

73. Slonim N, Atwal GS, Tkacik G, Bialek W: **Information-based clustering.** *Proc Natl Acad Sci U S A* 2005, **102:**18297-18302.

74. Beer MA, Tavazoie S: **Predicting gene expression from sequence.** *Cell* 2004, **117:**185-198.

75. Eisen MB, Spellman PT, Brown PO, Botstein D: **Cluster analysis and display of genome-wide expression patterns.** *Proc Natl Acad Sci U S A* 1998, **95:**14863-14868.

76. Zhou X, Kao MC, Wong WH: **Transitive functional annotation by shortest-path analysis of gene expression data.** *Proc Natl Acad Sci U S A* 2002, **99:**12783-12788.

77. Nariai N, Kolaczyk ED, Kasif S: **Probabilistic protein function prediction from heterogeneous genome-wide data.** *PLoS ONE* 2007, **2:**e337.

78. Shoemaker BA, Panchenko AR: **Deciphering protein-protein interactions. Part II. Computational methods to predict protein and domain interaction partners.** *PLoS Comput Biol* 2007, **3:**e43.

79. Lu LJ, Xia Y, Paccanaro A, Yu H, Gerstein M: **Assessing the limits of genomic data integration for predicting protein networks.** *Genome Res* 2005, **15:**945-953.

80. Date SV, Stoeckert CJ, Jr.: **Computational modeling of the Plasmodium falciparum interactome reveals protein function on a genome-wide scale.** *Genome Res* 2006, **16:**542-549.

81. Jansen R, Yu H, Greenbaum D, Kluger Y, Krogan NJ, Chung S, Emili A, Snyder M, Greenblatt JF, Gerstein M: **A Bayesian networks approach for predicting protein-protein interactions from genomic data.** *Science* 2003, **302:**449-453.

82. Marcotte EM, Pellegrini M, Ng HL, Rice DW, Yeates TO, Eisenberg D: **Detecting protein function and protein-protein interactions from genome sequences.** *Science* 1999, **285:**751-753.

83. Yu H, Luscombe NM, Lu HX, Zhu X, Xia Y, Han JD, Bertin N, Chung S, Vidal M, Gerstein M: **Annotation transfer between genomes: protein-protein interologs and protein-DNA regulogs.** *Genome Res* 2004, **14:**1107-1118.

84. Matthews LR, Vaglio P, Reboul J, Ge H, Davis BP, Garrels J, Vincent S, Vidal M: **Identification of potential interaction networks using sequence-based searches for conserved protein-protein interactions or "interologs".** *Genome Res* 2001, **11:**2120-2126.

85. Enright AJ, Iliopoulos I, Kyrpides NC, Ouzounis CA: **Protein interaction maps for complete genomes based on gene fusion events.** *Nature* 1999, **402:**86-90.

86. Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO: **Assigning protein functions by comparative genome analysis: protein phylogenetic profiles.** *Proc Natl Acad Sci U S A* 1999, **96:**4285-4288.

87. Date SV, Marcotte EM: **Discovery of uncharacterized cellular systems by genome-wide analysis of functional linkages.** *Nat Biotechnol* 2003, **21:**1055-1062.

88. Frearson JA, Wyatt PG, Gilbert IH, Fairlamb AH: **Target assessment for antiparasitic drug discovery.** *Trends Parasitol* 2007, **23:**589-595.

89. Simpson AG, Stevens JR, Lukes J: **The evolution and diversity of kinetoplastid flagellates.** *Trends Parasitol* 2006, **22:**168-174.

90. Stuart KD, Schnaufer A, Ernst NL, Panigrahi AK: **Complex management: RNA editing in trypanosomes.** *Trends Biochem Sci* 2005, **30:**97-105.

91. Lukes J, Hashimi H, Zikova A: **Unexplained complexity of the mitochondrial genome and transcriptome in kinetoplastid flagellates.** *Curr Genet* 2005, **48:**277-299.

92. Liu B, Liu Y, Motyka SA, Agbo EE, Englund PT: **Fellowship of the rings: the replication of kinetoplast DNA.** *Trends Parasitol* 2005, **21:**363-369.

93. Martinez-Calvillo S, Yan S, Nguyen D, Fox M, Stuart K, Myler PJ: **Transcription of Leishmania major Friedlin chromosome 1 initiates in both directions within a single region.** *Mol Cell* 2003, **11:**1291-1299.

94. Hannaert V, Bringaud F, Opperdoes FR, Michels PA: **Evolution of energy metabolism and its compartmentation in Kinetoplastida.** *Kinetoplastid Biol Dis* 2003, **2:**11.

95. Lee SH, Stephens JL, Paul KS, Englund PT: **Fatty acid synthesis by elongases in trypanosomes.** *Cell* 2006, **126:**691-699.

96. Lee SH, Stephens JL, Englund PT: **A fatty-acid synthesis mechanism specialized for parasitism.** *Nat Rev Microbiol* 2007, **5:**287-297.

97. Nagamune K, Nozaki T, Maeda Y, Ohishi K, Fukuma T, Hara T, Schwarz RT, Sutterlin C, Brun R, Riezman H, Kinoshita T: **Critical roles of glycosylphosphatidylinositol for Trypanosoma brucei.** *Proc Natl Acad Sci U S A* 2000, **97:**10336-10341.

98. Smith TK, Crossman A, Brimacombe JS, Ferguson MA: **Chemical validation of GPI biosynthesis as a drug target against African sleeping sickness.** *Embo J* 2004, **23:**4701-4708.

99. Ferguson MA: **Glycosylphosphatidylinositol biosynthesis validated as a drug target for African sleeping sickness.** *Proc Natl Acad Sci U S A* 2000, **97:**10673-10675.

100. Urbina JA, Docampo R: **Specific chemotherapy of Chagas disease: controversies and advances.** *Trends Parasitol* 2003, **19:**495-501.

101. Urbina JA: **Specific treatment of Chagas disease: current status and new developments.** *Curr Opin Infect Dis* 2001, **14:**733-741.

102. Urbina JA: **Chemotherapy of Chagas disease.** *Curr Pharm Des* 2002, **8:**287-295.

103. Beck JT, Ullman B: **Nutritional requirements of wild-type and folate transport-deficient Leishmania donovani for pterins and folates.** *Mol Biochem Parasitol* 1990, **43:**221-230.

104. Olliaro PL, Yuthavong Y: **An overview of chemotherapeutic targets for antimalarial drug discovery.** *Pharmacol Ther* 1999, **81:**91-110.

105. Senkovich O, Pal B, Schormann N, Chattopadhyay D: **Trypanosoma cruzi genome encodes a pteridine reductase 2 protein.** *Mol Biochem Parasitol* 2003, **127:**89-92.

106. Nare B, Luba J, Hardy LW, Beverley S: **New approaches to Leishmania chemotherapy: pteridine reductase 1 (PTR1) as a target and modulator of antifolate sensitivity.** *Parasitology* 1997, **114 Suppl:**S101-110.

107. Krauth-Siegel RL, Bauer H, Schirmer RH: **Dithiol proteins as guardians of the intracellular redox milieu in parasites: old and new drug targets in trypanosomes and malaria-causing plasmodia.** *Angew Chem Int Ed Engl* 2005, **44:**690-715.

108. Schmidt A, Krauth-Siegel RL: **Enzymes of the trypanothione metabolism as targets for antitrypanosomal drug development.** *Curr Top Med Chem* 2002, **2:**1239-1259.

109. Renslo AR, McKerrow JH: **Drug discovery and development for neglected parasitic diseases.** *Nat Chem Biol* 2006, **2:**701-710.

110. Opperdoes FR, Michels PA: **Enzymes of carbohydrate metabolism as potential drug targets.** *Int J Parasitol* 2001, **31:**482-490.

111. Roper JR, Guther ML, Macrae JI, Prescott AR, Hallyburton I, Acosta-Serrano A, Ferguson MA: **The suppression of galactose metabolism in procylic form Trypanosoma brucei causes cessation of cell growth and alters procyclin glycoprotein structure and copy number.** *J Biol Chem* 2005, **280:**19728-19736.

112. Roper JR, Guther ML, Milne KG, Ferguson MA: **Galactose metabolism is essential for the African sleeping sickness parasite Trypanosoma brucei.** *Proc Natl Acad Sci U S A* 2002, **99:**5884-5889.

113. Garami A, Mehlert A, Ilg T: **Glycosylation defects and virulence phenotypes of Leishmania mexicana phosphomannomutase and dolicholphosphate-mannose synthase gene deletion mutants.** *Mol Cell Biol* 2001, **21:**8168-8183.

114. Garami A, Ilg T: **Disruption of mannose activation in Leishmania mexicana: GDP-mannose pyrophosphorylase is required for virulence, but not for viability.** *Embo J* 2001, **20:**3657-3666.

115. Davis AJ, Perugini MA, Smith BJ, Stewart JD, Ilg T, Hodder AN, Handman E: **Properties of GDP-mannose pyrophosphorylase, a critical enzyme and drug target in Leishmania mexicana.** *J Biol Chem* 2004, **279:**12462-12468.

116. Szajnman SH, Garcia Linares GE, Li ZH, Jiang C, Galizzi M, Bontempi EJ, Ferella M, Moreno SN, Docampo R, Rodriguez JB: **Synthesis and biological evaluation of 2-alkylaminoethyl-1,1-bisphosphonic acids against**

Trypanosoma cruzi and Toxoplasma gondii targeting farnesyl diphosphate synthase. *Bioorg Med Chem* 2008, **16:**3283-3290.

117. Ferella M, Li ZH, Andersson B, Docampo R: **Farnesyl diphosphate synthase localizes to the cytoplasm of Trypanosoma cruzi and T. brucei.** *Exp Parasitol* 2008, **119:**308-312.

118. Buckner FS, Eastman RT, Yokoyama K, Gelb MH, Van Voorhis WC: **Protein farnesyl transferase inhibitors for the treatment of malaria and African trypanosomiasis.** *Curr Opin Investig Drugs* 2005, **6:**791-797.

119. Lakhdar-Ghazal F, Blonski C, Willson M, Michels P, Perie J: **Glycolysis and proteases as targets for the design of new anti-trypanosome drugs.** *Curr Top Med Chem* 2002, **2:**439-456.

120. Price HP, Menon MR, Panethymitaki C, Goulding D, McKean PG, Smith DF: **Myristoyl-CoA:protein N-myristoyltransferase, an essential enzyme and potential drug target in kinetoplastid parasites.** *J Biol Chem* 2003, **278:**7206-7214.

121. Heby O, Persson L, Rentala M: **Targeting the polyamine biosynthetic enzymes: a promising approach to therapy of African sleeping sickness, Chagas' disease, and leishmaniasis.** *Amino Acids* 2007, **33:**359-366.

122. Carter NS, Yates P, Arendt CS, Boitz JM, Ullman B: **Purine and pyrimidine metabolism in Leishmania.** *Adv Exp Med Biol* 2008, **625:**141-154.

123. Datta AK, Datta R, Sen B: **Antiparasitic chemotherapy: tinkering with the purine salvage pathway.** *Adv Exp Med Biol* 2008, **625:**116-132.

124. Doerig C: **Protein kinases as targets for anti-parasitic chemotherapy.** *Biochim Biophys Acta* 2004, **1697:**155-168.

125. Hammarton TC, Mottram JC, Doerig C: **The cell cycle of parasitic protozoa: potential for chemotherapeutic exploitation.** *Prog Cell Cycle Res* 2003, **5:**91-101.

126. Canduri F, Perez PC, Caceres RA, de Azevedo WF, Jr.: **Protein kinases as targets for antiparasitic chemotherapy drugs.** *Curr Drug Targets* 2007, **8:**389-398.

127. Balana-Fouce R, Redondo CM, Perez-Pertejo Y, Diaz-Gonzalez R, Reguera RM: **Targeting atypical trypanosomatid DNA topoisomerase I.** *Drug Discov Today* 2006, **11:**733-740.

128. Das BB, Ganguly A, Majumder HK: **DNA topoisomerases of Leishmania: the potential targets for anti-leishmanial therapy.** *Adv Exp Med Biol* 2008, **625:**103-115.

129. Das A, Dasgupta A, Sengupta T, Majumder HK: **Topoisomerases of kinetoplastid parasites as potential chemotherapeutic targets.** *Trends Parasitol* 2004, **20:**381-387.

130. Schnaufer A, Panigrahi AK, Panicucci B, Igo RP, Jr., Wirtz E, Salavati R, Stuart K: **An RNA ligase essential for RNA editing and survival of the bloodstream form of Trypanosoma brucei.** *Science* 2001, **291:**2159-2162.

131. Eisenberg D, Marcotte EM, Xenarios I, Yeates TO: **Protein function in the post-genomic era.** *Nature* 2000, **405:**823-826.

132. Shen J, Zhang J, Luo X, Zhu W, Yu K, Chen K, Li Y, Jiang H: **Predicting protein-protein interactions based only on sequences information.** *Proc Natl Acad Sci U S A* 2007, **104:**4337-4341.

133. Bock JR, Gough DA: **Predicting protein--protein interactions from primary structure.** *Bioinformatics* 2001, **17:**455-460.

134. Nanni L, Lumini A: **An ensemble of K-local hyperplanes for predicting protein-protein interactions.** *Bioinformatics* 2006, **22:**1207-1210.

135. Jansen R, Bussemaker HJ, Gerstein M: **Revisiting the codon adaptation index from a whole-genome perspective: analyzing the relationship between gene expression and codon occurrence in yeast using a variety of models.** *Nucleic Acids Res* 2003, **31:**2242-2251.

136. Daubin V, Perriere G: **G+C3 structuring along the genome: a common feature in prokaryotes.** *Mol Biol Evol* 2003, **20:**471-483.

137. Elf J, Nilsson D, Tenson T, Ehrenberg M: **Selective charging of tRNA isoacceptors explains patterns of codon usage.** *Science* 2003, **300:**1718-1722.

138. Dittmar KA, Sorensen MA, Elf J, Ehrenberg M, Pan T: **Selective charging of tRNA isoacceptors induced by amino-acid starvation.** *EMBO Rep* 2005, **6:**151-157.

139. Fraser HB, Hirsh AE, Wall DP, Eisen MB: **Coevolution of gene expression among interacting proteins.** *Proc Natl Acad Sci U S A* 2004, **101:**9033-9038.

140. Lithwick G, Margalit H: **Relative predicted protein levels of functionally associated proteins are conserved across organisms.** *Nucleic Acids Res* 2005, **33:**1051-1057.

141. Sharp PM, Li WH: **The codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications.** *Nucleic Acids Res* 1987, **15:**1281-1295.

142. Mewes HW, Amid C, Arnold R, Frishman D, Guldener U, Mannhaupt G, Munsterkotter M, Pagel P, Strack N, Stumpflen V, et al: **MIPS: analysis and annotation of proteins from whole genomes.** *Nucleic Acids Res* 2004, **32:**D41-44.

143. Kanehisa M, Goto S, Hattori M, Aoki-Kinoshita KF, Itoh M, Kawashima S, Katayama T, Araki M, Hirakawa M: **From genomics to chemical genomics: new developments in KEGG.** *Nucleic Acids Res* 2006, **34:**D354-357.

144. Arifuzzaman M, Maeda M, Itoh A, Nishikata K, Takita C, Saito R, Ara T, Nakahigashi K, Huang HC, Hirai A, et al: **Large-scale identification of protein-protein interaction of *Escherichia coli* K-12.** *Genome Res* 2006, **16:**686-691.

145. Musto H, Romero H, Zavala A, Jabbari K, Bernardi G: **Synonymous codon choices in the extremely GC-poor genome of *Plasmodium falciparum*: compositional constraints and translational selection.** *J Mol Evol* 1999, **49:**27-35.

146. Ubeda JM, Legare D, Raymond F, Ouameur AA, Boisvert S, Rigault P, Corbeil J, Tremblay MJ, Olivier M, Papadopoulou B, Ouellette M: **Modulation of gene expression in drug resistant Leishmania is associated with gene amplification, gene deletion and chromosome aneuploidy.** *Genome biology* 2008, **9:**R115.

147. Sharp PM, Tuohy TM, Mosurski KR: **Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed genes.** *Nucleic Acids Res* 1986, **14:**5125-5143.

148. Fuglsang A: **Estimating the "effective number of codons": the Wright way of determining codon homozygosity leads to superior estimates.** *Genetics* 2006, **172:**1301-1307.

149. Cabanas MJ, Vazquez D, Modolell J: **Dual interference of hygromycin B with ribosomal translocation and with aminoacyl-tRNA recognition.** *European journal of biochemistry / FEBS* 1978, **87:**21-27.

150. Krogan NJ, Cagney G, Yu H, Zhong G, Guo X, Ignatchenko A, Li J, Pu S, Datta N, Tikuisis AP, et al: **Global landscape of protein complexes in the yeast Saccharomyces cerevisiae.** *Nature* 2006, **440:**637-643.

151. Huh WK, Falvo JV, Gerke LC, Carroll AS, Howson RW, Weissman JS, O'Shea EK: **Global analysis of protein localization in budding yeast.** *Nature* 2003, **425:**686-691.

152. Gavin AC, Aloy P, Grandi P, Krause R, Boesche M, Marzioch M, Rau C, Jensen LJ, Bastuck S, Dumpelfeld B, et al: **Proteome survey reveals modularity of the yeast cell machinery.** *Nature* 2006, **440:**631-636.

153. Collins SR, Kemmeren P, Zhao XC, Greenblatt JF, Spencer F, Holstege FC, Weissman JS, Krogan NJ: **Toward a comprehensive atlas of the physical interactome of Saccharomyces cerevisiae.** *Mol Cell Proteomics* 2007, **6:**439-450.

154. Jao DL, Chen KY: **Tandem affinity purification revealed the hypusine-dependent binding of eukaryotic initiation factor 5A to the translating 80S ribosomal complex.** *J Cell Biochem* 2006, **97:**583-598.

155. Anand M, Chakraburtty K, Marton MJ, Hinnebusch AG, Kinzy TG: **Functional interactions between yeast translation eukaryotic elongation factor (eEF) 1A and eEF3.** *J Biol Chem* 2003, **278:**6985-6991.

156. Hebbar SK, Belcher SM, Perlman PS: **A maturase-encoding group IIA intron of yeast mitochondria self-splices in vitro.** *Nucleic Acids Res* 1992, **20:**1747-1754.

157. Belfort M: **Two for the price of one: a bifunctional intron-encoded DNA endonuclease-RNA maturase.** *Genes Dev* 2003, **17:**2860-2863.

158. Fleury D, Himanen K, Cnops G, Nelissen H, Boccardi TM, Maere S, Beemster GT, Neyt P, Anami S, Robles P, et al: **The Arabidopsis thaliana homolog of yeast BRE1 has a function in cell cycle regulation during early leaf and root growth.** *Plant Cell* 2007, **19:**417-432.

159. Jarrous N, Reiner R: **Human RNase P: a tRNA-processing enzyme and transcription factor.** *Nucleic Acids Res* 2007, **35:**3519-3524.

160. Reiner R, Ben-Asouli Y, Krilovetzky I, Jarrous N: **A role for the catalytic ribonucleoprotein RNase P in RNA polymerase III transcription.** *Genes Dev* 2006, **20:**1621-1635.

161. **Yeast PIC and PICT** [http://webpages.mcgill.ca/staff/Group2/rsalav/web/Supp18501006.htm]

162. Alvarez-Valin F, Tort JF, Bernardi G: **Nonrandom spatial distribution of synonymous substitutions in the GP63 gene from Leishmania.** *Genetics* 2000, **155:**1683-1692.

163. Stuart K, Gobright E, Jenni L, Milhausen M, Thomashow L, Agabian N: **The IsTaR 1 serodeme of Trypanosoma brucei: development of a new serodeme.** *The Journal of parasitology* 1984, **70:**747-754.

164. Queiroz R, Benz C, Fellenberg K, Hoheisel JD, Clayton C: **Transcriptome analysis of differentiating trypanosomes reveals the existence of multiple post-transcriptional regulons.** *BMC genomics* 2009, **10:**495.

165. Holzer TR, McMaster WR, Forney JD: **Expression profiling by whole-genome interspecies microarray hybridization reveals differential gene expression in procyclic promastigotes, lesion-derived amastigotes, and axenic amastigotes in Leishmania mexicana.** *Molecular and biochemical parasitology* 2006, **146:**198-218.

166. Kellis M, Patterson N, Endrizzi M, Birren B, Lander ES: **Sequencing and comparison of yeast species to identify genes and regulatory elements.** *Nature* 2003, **423:**241-254.

167. Greenbaum D, Jansen R, Gerstein M: **Analysis of mRNA expression and protein abundance data: an approach for the comparison of the enrichment of features in the cellular population of proteins and transcripts.** *Bioinformatics* 2002, **18:**585-596.

168. Elemento O, Slonim N, Tavazoie S: **A universal framework for regulatory element discovery across all genomes and data types.** *Mol Cell* 2007, **28:**337-350.

169. dos Santos G, Simmonds AJ, Krause HM: **A stem-loop structure in the wingless transcript defines a consensus motif for apical RNA transport.** *Development* 2008, **135:**133-143.

170. Merriam LC, Chess A: **cis-Regulatory elements within the odorant receptor coding region.** *Cell* 2007, **131:**844-846.

171. Lin X, Parsels LA, Voeller DM, Allegra CJ, Maley GF, Maley F, Chu E: **Characterization of a cis-acting regulatory element in the protein coding region of thymidylate synthase mRNA.** *Nucleic Acids Res* 2000, **28:**1381-1389.

172. Wenz P, Schwank S, Hoja U, Schuller HJ: **A downstream regulatory element located within the coding sequence mediates autoregulated expression of the yeast fatty acid synthase gene FAS2 by the FAS1 gene product.** *Nucleic Acids Res* 2001, **29:**4625-4632.

173. Sorensen MA, Elf J, Bouakaz E, Tenson T, Sanyal S, Bjork GR, Ehrenberg M: **Over expression of a tRNA(Leu) isoacceptor changes charging pattern of leucine tRNAs and reveals new codon reading.** *J Mol Biol* 2005, **354:**16-24.

174. Dittmar KA, Goodenbour JM, Pan T: **Tissue-specific differences in human transfer RNA expression.** *PLoS Genet* 2006, **2:**e221.

175. Begley U, Dyavaiah M, Patil A, Rooney JP, DiRenzo D, Young CM, Conklin DS, Zitomer RS, Begley TJ: **Trm9-catalyzed tRNA modifications link translation to the DNA damage response.** *Mol Cell* 2007, **28:**860-870.

176. Plotkin JB, Robins H, Levine AJ: **Tissue-specific codon usage and the expression of human genes.** *Proc Natl Acad Sci U S A* 2004, **101:**12588-12591.

177. Semon M, Lobry JR, Duret L: **No evidence for tissue-specific adaptation of synonymous codon usage in humans.** *Mol Biol Evol* 2006, **23:**523-529.

178. Chamary JV, Parmley JL, Hurst LD: **Hearing silence: non-neutral evolution at synonymous sites in mammals.** *Nat Rev Genet* 2006, **7:**98-108.

179. Lavner Y, Kotlar D: **Codon bias as a factor in regulating expression via translation rate in the human genome.** *Gene* 2005, **345:**127-138.

180. Comeron JM: **Selective and mutational patterns associated with gene expression in humans: influences on synonymous composition and intron presence.** *Genetics* 2004, **167:**1293-1304.

181. Kotlar D, Lavner Y: **The action of selection on codon bias in the human genome is related to frequency, complexity, and chronology of amino acids.** *BMC Genomics* 2006, **7:**67.

182. Chamary JV, Hurst LD: **Evidence for selection on synonymous mutations affecting stability of mRNA secondary structure in mammals.** *Genome Biol* 2005, **6:**R75.

183. Duan J, Wainwright MS, Comeron JM, Saitou N, Sanders AR, Gelernter J, Gejman PV: **Synonymous mutations in the human dopamine receptor D2 (DRD2) affect mRNA stability and synthesis of the receptor.** *Hum Mol Genet* 2003, **12:**205-216.

184. Capon F, Allen MH, Ameen M, Burden AD, Tillman D, Barker JN, Trembath RC: **A synonymous SNP of the corneodesmosin gene leads to increased mRNA stability and demonstrates association with psoriasis across diverse ethnic groups.** *Hum Mol Genet* 2004, **13:**2361-2368.

185. Cartegni L, Chew SL, Krainer AR: **Listening to silence and understanding nonsense: exonic mutations that affect splicing.** *Nat Rev Genet* 2002, **3:**285-298.

186. Sharp PM, Bailes E, Grocock RJ, Peden JF, Sockett RE: **Variation in the strength of selected codon usage bias among bacteria.** *Nucleic Acids Res* 2005, **33:**1141-1153.

187. Kudla G, Murray AW, Tollervey D, Plotkin JB: **Coding-sequence determinants of gene expression in Escherichia coli.** *Science* 2009, **324:**255-258.

188. Garel JP: **Functional adaptation of tRNA population.** *J Theor Biol* 1974, **43:**211-225.

189. Najafabadi HS, Salavati R: **Sequence-based prediction of protein-protein interactions by means of codon usage.** *Genome Biol* 2008, **9:**R87.

190. Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M: **The KEGG resource for deciphering the genome.** *Nucleic Acids Res* 2004, **32:**D277-280.

191. van Noort V, Snel B, Huynen MA: **Predicting gene function by conserved co-expression.** *Trends Genet* 2003, **19:**238-242.

192. Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, Block D, Zhang J, Soden R, Hayakawa M, Kreiman G, et al: **A gene atlas of the mouse and human protein-encoding transcriptomes.** *Proc Natl Acad Sci U S A* 2004, **101:**6062-6067.

193. Gasch AP, Spellman PT, Kao CM, Carmel-Harel O, Eisen MB, Storz G, Botstein D, Brown PO: **Genomic expression programs in the response of yeast cells to environmental changes.** *Mol Biol Cell* 2000, **11:**4241-4257.

194. Faith JJ, Hayete B, Thaden JT, Mogno I, Wierzbowski J, Cottarel G, Kasif S, Collins JJ, Gardner TS: **Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles.** *PLoS Biol* 2007, **5:**e8.

195. Kim SK, Lund J, Kiraly M, Duke K, Jiang M, Stuart JM, Eizinger A, Wylie BN, Davidson GS: **A gene expression map for Caenorhabditis elegans.** *Science* 2001, **293:**2087-2092.

196. Dunn RK, Kingston RE: **Gene regulation in the postgenomic era: technology takes the wheel.** *Mol Cell* 2007, **28:**708-714.

197. Kuhar I, van Putten JP, Zgur-Bertok D, Gaastra W, Jordi BJ: **Codon-usage based regulation of colicin K synthesis by the stress alarmone ppGpp.** *Mol Microbiol* 2001, **41:**207-216.

198. Baca AM, Hol WG: **Overcoming codon bias: a method for high-level overexpression of Plasmodium and other AT-rich parasite genes in Escherichia coli.** *Int J Parasitol* 2000, **30:**113-118.

199. Mituyama T, Yamada K, Hattori E, Okida H, Ono Y, Terai G, Yoshizawa A, Komori T, Asai K: **The Functional RNA Database 3.0: databases to support mining and annotation of functional RNAs.** *Nucleic Acids Res* 2009, **37:**D89-92.

200. Eddy SR: **Non-coding RNA genes and the modern RNA world.** *Nat Rev Genet* 2001, **2:**919-929.

201. Rabani M, Kertesz M, Segal E: **Computational prediction of RNA structural motifs involved in posttranscriptional regulatory processes.** *Proc Natl Acad Sci U S A* 2008, **105:**14885-14890.

202. Mallick B, Ghosh Z, Chakrabarti J: **MicroRNA switches in Trypanosoma brucei.** *Biochem Biophys Res Commun* 2008, **372:**459-463.

203. Pedersen JS, Bejerano G, Siepel A, Rosenbloom K, Lindblad-Toh K, Lander ES, Kent J, Miller W, Haussler D: **Identification and classification of conserved RNA secondary structures in the human genome.** *PLoS Comput Biol* 2006, **2:**e33.

204. Margulies EH, Blanchette M, Haussler D, Green ED: **Identification and characterization of multi-species conserved sequences.** *Genome Res* 2003, **13:**2507-2518.

205. Rivas E, Eddy SR: **Noncoding RNA gene detection using comparative sequence analysis.** *BMC Bioinformatics* 2001, **2:**8.

206. Grimson A, Srivastava M, Fahey B, Woodcroft BJ, Chiang HR, King N, Degnan BM, Rokhsar DS, Bartel DP: **Early origins and evolution of microRNAs and Piwi-interacting RNAs in animals.** *Nature* 2008, **455:**1193-1197.

207. Molnar A, Schwach F, Studholme DJ, Thuenemann EC, Baulcombe DC: **miRNAs control gene expression in the single-cell alga Chlamydomonas reinhardtii.** *Nature* 2007, **447:**1126-1129.

208. Koumandou VL, Natesan SK, Sergeenko T, Field MC: **The trypanosome transcriptome is remodelled during differentiation but displays limited responsiveness within life stages.** *BMC genomics* 2008, **9:**298.

209. Blattner J, Clayton CE: **The 3'-untranslated regions from the Trypanosoma brucei phosphoglycerate kinase-encoding genes mediate developmental regulation.** *Gene* 1995, **162:**153-156.

210. Cornish-Bowden A: **Nomenclature for incompletely specified bases in nucleic acid sequences: recommendations 1984.** *Nucleic Acids Res* 1985, **13:**3021-3030.

211. Kapotas N, Bellofatto V: **Differential response to RNA trans-splicing signals within the phosphoglycerate kinase gene cluster in Trypanosoma brucei.** *Nucleic Acids Res* 1993, **21:**4067-4072.

212. Siegel TN, Tan KS, Cross GA: **Systematic study of sequence motifs for RNA trans splicing in Trypanosoma brucei.** *Mol Cell Biol* 2005, **25:**9586-9594.

213. Schwartz S, Kent WJ, Smit A, Zhang Z, Baertsch R, Hardison RC, Haussler D, Miller W: **Human-mouse alignments with BLASTZ.** *Genome Res* 2003, **13:**103-107.

214. Washietl S, Hofacker IL, Stadler PF: **Fast and reliable prediction of noncoding RNAs.** *Proc Natl Acad Sci U S A* 2005, **102:**2454-2459.

215. Nam JW, Shin KR, Han J, Lee Y, Kim VN, Zhang BT: **Human microRNA prediction through a probabilistic co-learning model of sequence and structure.** *Nucleic Acids Res* 2005, **33:**3570-3581.

216. Kozlowski P, Starega-Roslan J, Legacz M, Magnus M, Krzyzosiak WJ: **Structures of MicroRNA Precursors.** In *Current Perspectives in microRNAs (miRNA).* Edited by Ying S-Y: Springer; 2008: 1-16

217. Lee D, Redfern O, Orengo C: **Predicting protein function from sequence and structure.** *Nat Rev Mol Cell Biol* 2007, **8:**995-1005.

218. Yooseph S, Sutton G, Rusch DB, Halpern AL, Williamson SJ, Remington K, Eisen JA, Heidelberg KB, Manning G, Li W, et al: **The Sorcerer II Global Ocean Sampling expedition: expanding the universe of protein families.** *PLoS Biol* 2007, **5:**e16.

219. Ellrott K, Jaroszewski L, Li W, Wooley JC, Godzik A: **Expansion of the protein repertoire in newly explored environments: human gut microbiome specific protein families.** *PLoS Comput Biol* 2010, **6:**e1000798.

220. Kannan S, Hauth AM, Burger G: **Function prediction of hypothetical proteins without sequence similarity to proteins of known function.** *Protein Pept Lett* 2008, **15:**1107-1116.

221. Neduva V, Russell RB: **Linear motifs: evolutionary interaction switches.** *FEBS Lett* 2005, **579:**3342-3345.

222. Hulo N, Bairoch A, Bulliard V, Cerutti L, De Castro E, Langendijk-Genevaux PS, Pagni M, Sigrist CJ: **The PROSITE database.** *Nucleic Acids Res* 2006, **34:**D227-230.

223. Gould CM, Diella F, Via A, Puntervoll P, Gemund C, Chabanis-Davidson S, Michael S, Sayadi A, Bryne JC, Chica C, et al: **ELM: the status of the 2010 eukaryotic linear motif resource.** *Nucleic Acids Res* 2010, **38:**D167-180.

224. Sigrist CJ, Cerutti L, de Castro E, Langendijk-Genevaux PS, Bulliard V, Bairoch A, Hulo N: **PROSITE, a protein domain database for functional characterization and annotation.** *Nucleic Acids Res* 2010, **38:**D161-166.

225. Tan SH, Hugo W, Sung WK, Ng SK: **A correlated motif approach for finding short linear motifs from protein interaction networks.** *BMC Bioinformatics* 2006, **7:**502.

226. Neduva V, Linding R, Su-Angrand I, Stark A, de Masi F, Gibson TJ, Lewis J, Serrano L, Russell RB: **Systematic discovery of new recognition peptides mediating protein interaction networks.** *PLoS Biol* 2005, **3:**e405.

227. Edwards RJ, Davey NE, Shields DC: **SLiMFinder: a probabilistic method for identifying over-represented, convergently evolved, short linear motifs in proteins.** *PLoS One* 2007, **2:**e967.

228. Lieber DS, Elemento O, Tavazoie S: **Large-scale discovery and characterization of protein regulatory motifs in eukaryotes.** *PLoS One* 2010, **5:**e14444.

229. Sarac OS, Atalay V, Cetin-Atalay R: **GOPred: GO molecular function prediction by combined classifiers.** *PLoS One* 2010, **5:**e12382.

230. Aslett M, Aurrecoechea C, Berriman M, Brestelli J, Brunk BP, Carrington M, Depledge DP, Fischer S, Gajria B, Gao X, et al: **TriTrypDB: a functional genomic resource for the Trypanosomatidae.** *Nucleic Acids Res* 2010, **38:**D457-462.

231. Hirokawa T, Boon-Chieng S, Mitaku S: **SOSUI: classification and secondary structure prediction system for membrane proteins.** *Bioinformatics* 1998, **14:**378-379.

232. Walker JE, Saraste M, Runswick MJ, Gay NJ: **Distantly related sequences in the alpha- and beta-subunits of ATP synthase, myosin, kinases and other ATP-requiring enzymes and a common nucleotide binding fold.** *EMBO J* 1982, **1:**945-951.

233. Rambaldi I, Kovacs EN, Featherstone MS: **A proline-rich transcriptional activation domain in murine HOXD-4 (HOX-4.2).** *Nucleic Acids Res* 1994, **22:**376-382.

234. Rawlings ND, Barrett AJ: **Evolutionary families of metallopeptidases.** *Methods Enzymol* 1995, **248:**183-228.

235. Bailey TL, Williams N, Misleh C, Li WW: **MEME: discovering and analyzing DNA and protein sequence motifs.** *Nucleic Acids Res* 2006, **34:**W369-373.

236. Degerman E, Belfrage P, Manganiello VC: **Structure, localization, and regulation of cGMP-inhibited phosphodiesterase (PDE3).** *J Biol Chem* 1997, **272:**6823-6826.

237. Desbois-Mouthon C, Cadoret A, Blivet-Van Eggelpoel MJ, Bertrand F, Cherqui G, Perret C, Capeau J: **Insulin and IGF-1 stimulate the beta-catenin**

pathway through two signalling cascades involving GSK-3beta inhibition and Ras activation. *Oncogene* 2001, **20:**252-259.

238. Yuan Y, Guo L, Shen L, Liu JS: **Predicting gene expression from sequence: a reexamination.** *PLoS Comput Biol* 2007, **3:**e243.

239. Krieger CJ, Zhang P, Mueller LA, Wang A, Paley S, Arnaud M, Pick J, Rhee SY, Karp PD: **MetaCyc: a multiorganism database of metabolic pathways and enzymes.** *Nucleic Acids Res* 2004, **32:**D438-442.

240. Peacock CS, Seeger K, Harris D, Murphy L, Ruiz JC, Quail MA, Peters N, Adlem E, Tivey A, Aslett M, et al: **Comparative genomic analysis of three *Leishmania* species that cause diverse human disease.** *Nat Genet* 2007, **39:**839-847.

241. Sharan R, Ulitsky I, Shamir R: **Network-based prediction of protein function.** *Mol Syst Biol* 2007, **3:**88.

242. Slonim N, Atwal GS, Tkacik G, Bialek W: **Information-based clustering.** *Proc Natl Acad Sci U S A* 2005, **102:**18297-18302.

243. Eisen MB, Spellman PT, Brown PO, Botstein D: **Cluster analysis and display of genome-wide expression patterns.** *Proc Natl Acad Sci U S A* 1998, **95:**14863-14868.

244. Zhou X, Kao MC, Wong WH: **Transitive functional annotation by shortest-path analysis of gene expression data.** *Proc Natl Acad Sci U S A* 2002, **99:**12783-12788.

245. Shoemaker BA, Panchenko AR: **Deciphering protein-protein interactions. Part II. Computational methods to predict protein and domain interaction partners.** *PLoS Comput Biol* 2007, **3:**e43.

246. Lu LJ, Xia Y, Paccanaro A, Yu H, Gerstein M: **Assessing the limits of genomic data integration for predicting protein networks.** *Genome Res* 2005, **15:**945-953.

247. Date SV, Stoeckert CJ, Jr.: **Computational modeling of the *Plasmodium falciparum* interactome reveals protein function on a genome-wide scale.** *Genome Res* 2006, **16:**542-549.

248. Yu H, Luscombe NM, Lu HX, Zhu X, Xia Y, Han JD, Bertin N, Chung S, Vidal M, Gerstein M: **Annotation transfer between genomes: protein-protein interologs and protein-DNA regulogs.** *Genome Res* 2004, **14:**1107-1118.

249. Matthews LR, Vaglio P, Reboul J, Ge H, Davis BP, Garrels J, Vincent S, Vidal M: **Identification of potential interaction networks using sequence-based searches for conserved protein-protein interactions or "interologs".** *Genome Res* 2001, **11:**2120-2126.

250. Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO: **Assigning protein functions by comparative genome analysis: protein phylogenetic profiles.** *Proc Natl Acad Sci U S A* 1999, **96:**4285-4288.

251. Date SV, Marcotte EM: **Discovery of uncharacterized cellular systems by genome-wide analysis of functional linkages.** *Nat Biotechnol* 2003, **21:**1055-1062.

252. Shateri Najafabadi H, Salavati R: **Sequence-based prediction of protein-protein interactions by means of codon usage.** *Genome Biol* 2008, **9:**R87.

253. Najafabadi HS, Goodarzi H, Salavati R: **Universal function-specificity of codon usage.** *Nucleic Acids Res* 2009, **doi:10.1093/nar/gkp792**.

254. Benz C, Nilsson D, Andersson B, Clayton C, Guilbride DL: **Messenger RNA processing sites in *Trypanosoma brucei*.** *Mol Biochem Parasitol* 2005, **143:**125-134.

255. Hotz HR, Hartmann C, Huober K, Hug M, Clayton C: **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* 1997, **25:**3017-3026.

256. Irmer H, Clayton C: **Degradation of the unstable EP1 mRNA in *Trypanosoma brucei* involves initial destruction of the 3'-untranslated region.** *Nucleic Acids Res* 2001, **29:**4707-4715.

257. Schurch N, Furger A, Kurath U, Roditi I: **Contributions of the procyclin 3' untranslated region and coding region to the regulation of expression in bloodstream forms of *Trypanosoma brucei*.** *Mol Biochem Parasitol* 1997, **89:**109-121.

258. Furger A, Schurch N, Kurath U, Roditi I: **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* 1997, **17:**4372-4380.

259. Hehl A, Vassella E, Braun R, Roditi I: **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* 1994, **91:**370-374.

260. Vassella E, Probst M, Schneider A, Studer E, Renggli CK, Roditi I: **Expression of a major surface protein of *Trypanosoma brucei* insect forms is controlled by the activity of mitochondrial enzymes.** *Mol Biol Cell* 2004, **15:**3986-3993.

261. Vassella E, Den Abbeele JV, Butikofer P, Renggli CK, Furger A, Brun R, Roditi I: **A major surface glycoprotein of *Trypanosoma brucei* is expressed transiently during development and can be regulated post-transcriptionally by glycerol or hypoxia.** *Genes Dev* 2000, **14:**615-626.

262. Quijada L, Guerra-Giraldez C, Drozdz M, Hartmann C, Irmer H, Ben-Dov C, Cristodero M, Ding M, Clayton C: **Expression of the human RNA-binding protein HuR in *Trypanosoma brucei* increases the abundance of mRNAs containing AU-rich regulatory elements.** *Nucleic Acids Res* 2002, **30:**4414-4424.

263. Webb H, Burns R, Ellis L, Kimblin N, Carrington M: **Developmentally regulated instability of the GPI-PLC mRNA is dependent on a short-lived protein factor.** *Nucleic Acids Res* 2005, **33:**1503-1512.

264. Mishra KK, Holzer TR, Moore LL, LeBowitz JH: **A negative regulatory element controls mRNA abundance of the *Leishmania mexicana* Paraflagellar rod gene PFR2.** *Euk Cell* 2003, **2:**1009-1017.

265. Purdy JE, Donelson JE, Wilson ME: **Regulation of genes encoding the major surface protease of *Leishmania chagasi* via mRNA stability.** *Mol Biochem Parasitol* 2005, **142:**88-97.

266. Teixeira SM, Kirchhoff LV, Donelson JE: **Post-transcriptional elements regulating expression of mRNAs from the amastin/tuzin gene cluster of *Trypanosoma cruzi*.** *J Biol Chem* 1995, **270:**22586-22594.

267. Coughlin BC, Teixeira SM, Kirchhoff LV, Donelson JE: **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* 2000, **275:**12051-12060.

268. Mayho M, Fenn K, Craddy P, Crosthwaite S, Matthews K: **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* 2006, **34:**5312-5324.

269. D'Orso I, Frasch AC: **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* 2002, **277:**50520-50528.

270. Di Noia JM, D'Orso I, Sanchez DO, Frasch AC: **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* 2000, **275:**10218-10227.

271. Bringaud F, Muller M, Cerqueira GC, Smith M, Rochette A, El-Sayed NM, Papadopoulou B, Ghedin E: **Members of a large retroposon family are determinants of post-transcriptional gene expression in *Leishmania*.** *PLoS Pathog* 2007, **3:**1291-1307.

272. McNicoll F, Muller M, Cloutier S, Boilard N, Rochette A, Dube M, Papadopoulou B: **Distinct 3'-untranslated region elements regulate stage-specific mRNA accumulation and translation in *Leishmania*.** *J Biol Chem* 2005, **280:**35238-35246.

273. Boucher N, Wu Y, Dumas C, Dube M, Sereno D, Breton M, Papadopoulou B: **A common mechanism of stage-regulated gene expression in *Leishmania* mediated by a conserved 3'-untranslated region element.** *J Biol Chem* 2002, **277:**19511-19520.

274. Engstler M, Boshart M: **Cold shock and regulation of surface protein trafficking convey sensitization to inducers of stage differentiation in *Trypanosoma brucei*.** *Genes Dev* 2004, **18:**2798-2811.

275. Colasante C, Robles A, Li CH, Schwede A, Benz C, Voncken F, Guilbride DL, Clayton C: **Regulated expression of glycosomal phosphoglycerate kinase in *Trypanosoma brucei*.** *Mol Biochem Parasitol* 2007, **151:**193-204.

276. Zilka A, Garlapati S, Dahan E, Yaolsky V, Shapira M: **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* 2001, **276:**47922-47929.

277. Quijada L, Soto M, Alonso C, Requena JM: **Identification of a putative regulatory element in the 3'-untranslated region that controls expression of HSP70 in *Leishmania infantum*.** *Mol Biochem Parasitol* 2000, **110:**79-91.

278. Murray A, Fu C, Habibi G, McMaster WR: **Regions in the 3' untranslated region confer stage-specific expression to the *Leishmania mexicana* a600-4 gene.** *Mol Biochem Parasitol* 2007, **153:**125-132.

279. Elemento O, Slonim N, Tavazoie S: **A universal framework for regulatory element discovery across all genomes and data types.** *Mol Cell* 2007, **28:**337-350.

280. Chung DW, Ponts N, Cervantes S, Le Roch KG: **Post-translational modifications in Plasmodium: more than you think!** *Mol Biochem Parasitol* 2009, **168:**123-134.

281. Diella F, Haslam N, Chica C, Budd A, Michael S, Brown NP, Trave G, Gibson TJ: **Understanding eukaryotic linear motifs and their role in cell signaling and regulation.** *Front Biosci* 2008, **13:**6580-6603.

282. DeRisi JL, Iyer VR, Brown PO: **Exploring the metabolic and genetic control of gene expression on a genomic scale.** *Science* 1997, **278:**680-686.

283. Stuart JM, Segal E, Koller D, Kim SK: **A gene-coexpression network for global discovery of conserved genetic modules.** *Science* 2003, **302:**249-255.

284. Kanehisa M, Goto S: **KEGG: kyoto encyclopedia of genes and genomes.** *Nucleic Acids Res* 2000, **28:**27-30.

285. Aslett M, Aurrecoechea C, Berriman M, Brestelli J, Brunk BP, Carrington M, Depledge DP, Fischer S, Gajria B, Gao X, et al: **TriTrypDB: a functional genomic resource for the Trypanosomatidae.** *Nucleic Acids Res*, **38:**D457-462.

286. Ubeda JM, Legare D, Raymond F, Ouameur AA, Boisvert S, Rigault P, Corbeil J, Tremblay MJ, Olivier M, Papadopoulou B, Ouellette M: **Modulation of gene**

expression in drug resistant Leishmania is associated with gene amplification, gene deletion and chromosome aneuploidy. *Genome Biol* 2008, **9:**R115.

287. Rochette A, Raymond F, Corbeil J, Ouellette M, Papadopoulou B: **Whole-genome comparative RNA expression profiling of axenic and intracellular amastigote forms of Leishmania infantum.** *Mol Biochem Parasitol* 2009, **165:**32-47.

288. Leprohon P, Legare D, Raymond F, Madore E, Hardiman G, Corbeil J, Ouellette M: **Gene expression modulation is associated with gene amplification, supernumerary chromosomes and chromosome loss in antimony-resistant Leishmania infantum.** *Nucleic Acids Res* 2009, **37:**1387-1399.

289. Zikova A, Panigrahi AK, Uboldi AD, Dalley RA, Handman E, Stuart K: **Structural and functional association of Trypanosoma brucei MIX protein with cytochrome c oxidase complex.** *Eukaryot Cell* 2008, **7:**1994-2003.

290. Zikova A, Schnaufer A, Dalley RA, Panigrahi AK, Stuart KD: **The F(0)F(1)-ATP synthase complex contains novel subunits and is essential for procyclic Trypanosoma brucei.** *PLoS Pathog* 2009, **5:**e1000436.

291. Bouvet P, Diaz JJ, Kindbeiter K, Madjar JJ, Amalric F: **Nucleolin interacts with several ribosomal proteins through its RGG domain.** *J Biol Chem* 1998, **273:**19025-19029.

292. Ramani AK, Li Z, Hart GT, Carlson MW, Boutz DR, Marcotte EM: **A map of human protein interactions derived from co-expression of human mRNAs and their orthologs.** *Mol Syst Biol* 2008, **4:**180.

293. Panigrahi AK, Ogata Y, Zikova A, Anupama A, Dalley RA, Acestor N, Myler PJ, Stuart KD: **A comprehensive analysis of Trypanosoma brucei mitochondrial proteome.** *Proteomics* 2009, **9:**434-450.

294. Acestor N, Panigrahi AK, Ogata Y, Anupama A, Stuart KD: **Protein composition of Trypanosoma brucei mitochondrial membranes.** *Proteomics* 2009, **9:**5497-5508.

295. Nayak RR, Kearns M, Spielman RS, Cheung VG: **Coexpression network based on natural variation in human gene expression reveals gene interactions and functions.** *Genome Res* 2009, **19:**1953-1962.

296. Czarna M, Jarmuszkiewicz W: **Activation of alternative oxidase and uncoupling protein lowers hydrogen peroxide formation in amoeba Acanthamoeba castellanii mitochondria.** *FEBS Lett* 2005, **579:**3136-3140.

297. Lecordier L, Devaux S, Uzureau P, Dierick JF, Walgraffe D, Poelvoorde P, Pays E, Vanhamme L: **Characterization of a TFIIH homologue from Trypanosoma brucei.** *Mol Microbiol* 2007, **64:**1164-1181.

298. Lee JH, Jung HS, Gunzl A: **Transcriptionally active TFIIH of the early-diverged eukaryote Trypanosoma brucei harbors two novel core subunits but not a cyclin-activating kinase complex.** *Nucleic Acids Res* 2009, **37:**3811-3820.

299. Lemtiri-Chlieh F, MacRobbie EA, Brearley CA: **Inositol hexakisphosphate is a physiological signal regulating the K+-inward rectifying conductance in guard cells.** *Proc Natl Acad Sci U S A* 2000, **97:**8687-8692.

300. Bian J, Cui J, McDonald TV: **HERG K(+) channel activity is regulated by changes in phosphatidyl inositol 4,5-bisphosphate.** *Circ Res* 2001, **89:**1168-1176.

301. Filosa JA, Bonev AD, Straub SV, Meredith AL, Wilkerson MK, Aldrich RW, Nelson MT: **Local potassium signaling couples neuronal activity to vasodilation in the brain.** *Nat Neurosci* 2006, **9:**1397-1403.

302. Colasante C, Ellis M, Ruppert T, Voncken F: **Comparative proteomics of glycosomes from bloodstream form and procyclic culture form Trypanosoma brucei brucei.** *Proteomics* 2006, **6:**3275-3293.

303. Andres JL, Johansen JW, Maller JL: **Identification of protein phosphatases 1 and 2B as ribosomal protein S6 phosphatases in vitro and in vivo.** *J Biol Chem* 1987, **262:**14389-14393.

304. Tavazoie S, Hughes JD, Campbell MJ, Cho RJ, Church GM: **Systematic determination of genetic network architecture.** *Nature genetics* 1999, **22:**281-285.

305. Cheadle C, Fan J, Cho-Chung YS, Werner T, Ray J, Do L, Gorospe M, Becker KG: **Control of gene expression during T cell activation: alternate regulation of mRNA transcription and mRNA stability.** *BMC genomics* 2005, **6:**75.

306. Perez-Ortin JE, Alepuz PM, Moreno J: **Genomics and gene transcription kinetics in yeast.** *Trends in genetics : TIG* 2007, **23:**250-257.

307. Bolognani F, Perrone-Bizzozero NI: **RNA-protein interactions and control of mRNA stability in neurons.** *Journal of neuroscience research* 2008, **86:**481-489.

308. Hughes TA: **Regulation of gene expression by alternative untranslated regions.** *Trends in genetics : TIG* 2006, **22:**119-122.

309. Anantharaman V, Koonin EV, Aravind L: **Comparative genomics and evolution of proteins involved in RNA metabolism.** *Nucleic acids research* 2002, **30:**1427-1464.

310. Kertesz M, Iovino N, Unnerstall U, Gaul U, Segal E: **The role of site accessibility in microRNA target recognition.** *Nature genetics* 2007, **39:**1278-1284.

311. Kertesz M, Wan Y, Mazor E, Rinn JL, Nutter RC, Chang HY, Segal E: **Genome-wide measurement of RNA secondary structure in yeast.** *Nature* 2010, **467:**103-107.

312. Rabani M, Kertesz M, Segal E: **Computational prediction of RNA structural motifs involved in posttranscriptional regulatory processes.** *Proceedings of the National Academy of Sciences of the United States of America* 2008, **105:**14885-14890.

313. Foat BC, Stormo GD: **Discovering structural cis-regulatory elements by modeling the behaviors of mRNAs.** *Molecular systems biology* 2009, **5:**268.

314. Pavesi G, Mauri G, Stefani M, Pesole G: **RNAProfile: an algorithm for finding conserved secondary structure motifs in unaligned RNA sequences.** *Nucleic acids research* 2004, **32:**3258-3269.

315. Xie X, Lu J, Kulbokas EJ, Golub TR, Mootha V, Lindblad-Toh K, Lander ES, Kellis M: **Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals.** *Nature* 2005, **434:**338-345.

316. Hofacker IL, Fekete M, Flamm C, Huynen MA, Rauscher S, Stolorz PE, Stadler PF: **Automatic detection of conserved RNA structure elements in complete RNA virus genomes.** *Nucleic acids research* 1998, **26:**3825-3836.

317. Gorodkin J, Heyer LJ, Stormo GD: **Finding the most significant common sequence and structure motifs in a set of RNA sequences.** *Nucleic acids research* 1997, **25:**3724-3732.

318. Washietl S, Hofacker IL, Stadler PF: **Fast and reliable prediction of noncoding RNAs.** *Proceedings of the National Academy of Sciences of the United States of America* 2005, **102:**2454-2459.

319. Pritsker M, Liu YC, Beer MA, Tavazoie S: **Whole-genome discovery of transcription factor binding sites by network-level conservation.** *Genome research* 2004, **14:**99-108.

320. Chan CS, Elemento O, Tavazoie S: **Revealing posttranscriptional regulatory elements through network-level conservation.** *PLoS computational biology* 2005, **1:**e69.

321. Haile S, Papadopoulou B: **Developmental regulation of gene expression in trypanosomatid parasitic protozoa.** *Current opinion in microbiology* 2007, **10:**569-577.

322. Clayton C, Shapira M: **Post-transcriptional regulation of gene expression in trypanosomes and leishmanias.** *Molecular and biochemical parasitology* 2007, **156:**93-101.

323. Babu MM, Luscombe NM, Aravind L, Gerstein M, Teichmann SA: **Structure and evolution of transcriptional regulatory networks.** *Current opinion in structural biology* 2004, **14:**283-291.

324. Bhandari D, Guha K, Bhaduri N, Saha P: **Ubiquitination of mRNA cycling sequence binding protein from Leishmania donovani (LdCSBP) modulates the RNA endonuclease activity of its Smr domain.** *FEBS letters* 2011, **585:**809-813.

325. Williams AS, Marzluff WF: **The sequence of the stem and flanking sequences at the 3' end of histone mRNA are critical determinants for the binding of the stem-loop binding protein.** *Nucleic acids research* 1995, **23:**654-662.

326. Gerber AP, Luschnig S, Krasnow MA, Brown PO, Herschlag D: **Genome-wide identification of mRNAs associated with the translational regulator PUMILIO in Drosophila melanogaster.** *Proceedings of the National Academy of Sciences of the United States of America* 2006, **103:**4487-4492.

327. Archer SK, Inchaustegui D, Queiroz R, Clayton C: **The cell cycle regulated transcriptome of Trypanosoma brucei.** *PloS one* 2011, **6:**e18425.

328. Barreau C, Paillard L, Osborne HB: **AU-rich elements and associated factors: are there unifying principles?** *Nucleic acids research* 2005, **33:**7138-7150.

329. Peng SS, Chen CY, Xu N, Shyu AB: **RNA stabilization by the AU-rich element binding protein, HuR, an ELAV protein.** *The EMBO journal* 1998, **17:**3461-3470.

330. Ford LP, Watson J, Keene JD, Wilusz J: **ELAV proteins stabilize deadenylated intermediates in a novel in vitro mRNA deadenylation/degradation system.** *Genes & development* 1999, **13:**188-201.

331. Fischer N, Weis K: **The DEAD box protein Dhh1 stimulates the decapping enzyme Dcp1.** *The EMBO journal* 2002, **21:**2788-2797.

332. Ma WJ, Chung S, Furneaux H: **The Elav-like proteins bind to AU-rich elements and to the poly(A) tail of mRNA.** *Nucleic acids research* 1997, **25:**3564-3569.

333. Quijada L, Guerra-Giraldez C, Drozdz M, Hartmann C, Irmer H, Ben-Dov C, Cristodero M, Ding M, Clayton C: **Expression of the human RNA-binding protein HuR in Trypanosoma brucei increases the abundance of mRNAs containing AU-rich regulatory elements.** *Nucleic acids research* 2002, **30:**4414-4424.

334. Mukherjee D, Gao M, O'Connor JP, Raijmakers R, Pruijn G, Lutz CS, Wilusz J: **The mammalian exosome mediates the efficient degradation of mRNAs that contain AU-rich elements.** *The EMBO journal* 2002, **21:**165-174.

335. Anderson JR, Mukherjee D, Muthukumaraswamy K, Moraes KC, Wilusz CJ, Wilusz J: **Sequence-specific RNA binding mediated by the RNase PH domain of components of the exosome.** *RNA* 2006, **12:**1810-1816.

336. Haile S, Estevez AM, Clayton C: **A role for the exosome in the in vivo degradation of unstable mRNAs.** *RNA* 2003, **9:**1491-1501.

337. Roy Chowdhury A, Bakshi R, Wang J, Yildirir G, Liu B, Pappas-Brown V, Tolun G, Griffith JD, Shapiro TA, Jensen RE, Englund PT: **The killing of African trypanosomes by ethidium bromide.** *PLoS pathogens* 2010, **6:**e1001226.

338. Kaminsky R, Zweygarth E: **The effect of verapamil alone and in combination with trypanocides on multidrug-resistant Trypanosoma brucei brucei.** *Acta tropica* 1991, **49:**215-225.

339. Aslett M, Aurrecoechea C, Berriman M, Brestelli J, Brunk BP, Carrington M, Depledge DP, Fischer S, Gajria B, Gao X, et al: **TriTrypDB: a functional genomic resource for the Trypanosomatidae.** *Nucleic acids research* 2010, **38:**D457-462.

340. Chen F, Mackey AJ, Stoeckert CJ, Jr., Roos DS: **OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups.** *Nucleic acids research* 2006, **34:**D363-368.

341. Rochette A, Raymond F, Ubeda JM, Smith M, Messier N, Boisvert S, Rigault P, Corbeil J, Ouellette M, Papadopoulou B: **Genome-wide gene expression profiling analysis of Leishmania major and Leishmania infantum developmental stages reveals substantial differences between the two species.** *BMC genomics* 2008, **9:**255.

342. Sayers EW, Barrett T, Benson DA, Bolton E, Bryant SH, Canese K, Chetvernin V, Church DM, DiCuccio M, Federhen S, et al: **Database resources of the National Center for Biotechnology Information.** *Nucleic Acids Res* 2011, **39:**D38-51.

343. **The Universal Protein Resource (UniProt) in 2010.** *Nucleic Acids Res* 2010, **38:**D142-148.

344. Boseley S: **New treatments raise hope of cutting sleeping sickness deaths.** In *The Guardian*. London; 2009.

345. Alsford S, Turner DJ, Obado SO, Sanchez-Flores A, Glover L, Berriman M, Hertz-Fowler C, Horn D: **High-throughput phenotyping using parallel sequencing of RNA interference targets in the African trypanosome.** *Genome Res* 2011, **21:**915-924.

346. Ghazalpour A, Doss S, Zhang B, Wang S, Plaisier C, Castellanos R, Brozell A, Schadt EE, Drake TA, Lusis AJ, Horvath S: **Integrating genetic and network analysis to characterize genes related to mouse weight.** *PLoS genetics* 2006, **2:**e130.

347. Hughes TR, Marton MJ, Jones AR, Roberts CJ, Stoughton R, Armour CD, Bennett HA, Coffey E, Dai H, He YD, et al: **Functional discovery via a compendium of expression profiles.** *Cell* 2000, **102:**109-126.

348. Roguev A, Bandyopadhyay S, Zofall M, Zhang K, Fischer T, Collins SR, Qu H, Shales M, Park HO, Hayles J, et al: **Conservation and rewiring of functional modules revealed by an epistasis map in fission yeast.** *Science* 2008, **322:**405-410.

349. Bandyopadhyay S, Mehta M, Kuo D, Sung MK, Chuang R, Jaehnig EJ, Bodenmiller B, Licon K, Copeland W, Shales M, et al: **Rewiring of genetic networks in response to DNA damage.** *Science* 2010, **330:**1385-1389.

350. Mani J, Guttinger A, Schimanski B, Heller M, Acosta-Serrano A, Pescher P, Spath G, Roditi I: **Alba-Domain Proteins of Trypanosoma brucei Are Cytoplasmic RNA-Binding Proteins That Interact with the Translation Machinery.** *PLoS ONE* 2011, **6:**e22463.

351. Archer SK, Luu VD, de Queiroz RA, Brems S, Clayton C: **Trypanosoma brucei PUF9 regulates mRNAs for proteins involved in replicative processes over the cell cycle.** *PLoS pathogens* 2009, **5:**e1000565.

352. Keene JD, Komisarow JM, Friedersdorf MB: **RIP-Chip: the isolation and identification of mRNAs, microRNAs and protein components of ribonucleoprotein complexes from cell extracts.** *Nature protocols* 2006, **1:**302-307.

353. Hafner M, Landthaler M, Burger L, Khorshid M, Hausser J, Berninger P, Rothballer A, Ascano M, Jr., Jungkamp AC, Munschauer M, et al: **Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP.** *Cell* 2010, **141:**129-141.