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

**Supplementary Methods**

Hamed Shateri Najafabadi$^{1,2}$ and Reza Salavati$^{1,2,3,*}$

1 Institute of Parasitology, McGill University, 21,111 Lakeshore Road, Ste. Anne de Bellevue, Quebec H9X3V9, Canada.

2 McGill Centre for Bioinformatics, McGill University, Duff Medical Building, 3775 University Street, Montreal, Quebec H3A2B4, Canada.

3 Department of Biochemistry, McGill University, McIntyre Medical Building, 3655 Promenade Sir William Osler, Montreal, Quebec H3G1Y6, Canada.

* Corresponding author:
Reza Salavati
Institute of Parasitology, Room A-208
21111 Lakeshore Road, Ste. Anne de Bellevue, Quebec H9X 3V9, Canada
Tel: 514-398-7721
Fax: 514-398-7857
Email: reza.salavati@mcgill.ca
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: \( x = (x' - μ)/σ \), 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:

![Normalized expression values across different experiments](image-url)
The coexpression network $G^E_\theta$ is the set of all gene pairs whose Pearson correlation coefficients, according to the experiment set $E$, are at least $\theta$:

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

where $\rho$ is the Pearson correlation coefficient function. The set of nodes in the coexpression network $G^E_\theta$ is denoted as $N^E_\theta$:

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

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

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

Thus, $f^E_\theta$ indicates what fraction of all genes the network $G^E_\theta$ 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 $G^E_{\theta}$ is defined as:

$$p^E_{\theta} = \frac{|G^E_{\theta} \cap I|}{(|G^E_{\theta} \cap I| + |G^E_{\theta} \cap I'|)}$$

Therefore, $p^E_{\theta}$ estimates the fraction of gene pairs in $G^E_{\theta}$ that are functionally related. We used the area under the curve (AUC) for $p^E_{\theta}(0)$ vs. $f^E_{\theta}(0)$ 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. [1], $E_Q$ which is the set of eight experiments from ref. [2], $E_J$ which is the set of five experiments from ref. [3], $E_{KQJ} = E_K + E_Q + E_J$, and $E \subseteq E_{KQJ}$. $E$ is chosen so as to result in the maximum $A^E$:

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

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

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

Using each of the experiment sets $E_{KQJ}$ and $E$, we defined a coexpression network by selecting the minimum value for cutoff $\theta$ that could result in $p_\theta \geq 0.75$ (i.e. precision of at least 75%). The selected value of $\theta$ for $E_{KQJ}$ was 0.94 and for $E$ was 0.957. The resulting networks are referred to in the paper as CoExp$^1\text{Tbr}$ and CoExp$^2\text{Tbr}$, respectively.
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 [4,5,6]. The conserved coexpression network $G^{E,E'}_{\theta,\theta'}$ is the set of all gene pairs that are coexpressed according to both experiment sets $E=K_{QJ}$ (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 $\rho$ 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 $\theta$ and $\theta'$ values, 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 $\theta$ and $\theta'$ values, we performed a leave-one-out cross-validation, in which each time one gene pair $(l,k)$ was left out, the best $\theta$ and $\theta'$ values were determined using the remaining gene pairs, and the left out gene pair was evaluated using these values. If $\rho(X^E_{i'},X^E_{l})\geq \theta$ and $\rho(X^{E'}_{i'},X^{E'}_{k})\geq \theta'$, the pair $(l,k)$ was added to the cross-validation network $G^x$:

1. Set $G^x=\emptyset$
2. For all $\{(l,k)|i\in S', j\in S', (l,k)\in I\cup I'\}$
3. \quad If $(l,k)\in I$ then $I=I\setminus\{(l,k)\}$
4. \quad If $(l,k)\in I'$ then $I'=I'\setminus\{(l,k)\}$
5. \quad Find the values for $\theta$ and $\theta'$ using the new $I$ and $I'$
6. \quad If $\rho(X^E_{i'},X^E_{l})\geq \theta$ and $\rho(X^{E'}_{i'},X^{E'}_{k})\geq \theta'$ then $G^x= G^x+\{(l,k)\}$
7. \quad Restore $I$ and $I'$
8. 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 $\theta$ and $\theta'$ did not over-train them.

**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 $G$, $C_\alpha \subseteq N$ is the set of nodes that are connected to the node $\alpha$ (excluding the node $\alpha$ itself), and $M \subseteq N$ is the set of nodes that have the particular function $f^M$ according to KEGG, again excluding the node $\alpha$ itself:

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

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 $\alpha$ is considered associated with $M$ and, thus, with function $f^M$. Since node $\alpha$ 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.

**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 [7] 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 [8] 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](http://webpages.mcgill.ca/staff/Group2/rsalav/web/Suppl/20100109/index.htm).
References


