PLOS | ONE

RESEARCH ARTICLE

Systems Biology Approach to Model the Life Cycle of *Trypanosoma cruzi*

Alejandra Carrea, Luis Diambra*

Centro Regional de Estudios Genómicos, Universidad Nacional de La Plata, La Plata, Argentina

* ldiambra@gmail.com

Abstract

Due to recent advances in reprogramming cell phenotypes, many efforts have been dedicated to developing reverse engineering procedures for the identification of gene regulatory networks that emulate dynamical properties associated with the cell fates of a given biological system. In this work, we propose a systems biology approach for the reconstruction of the gene regulatory network underlying the dynamics of the *Trypanosoma cruzi*'s life cycle. By means of an optimisation procedure, we embedded the steady state maintenance, and the known phenotypic transitions between these steady states in response to environmental cues, into the dynamics of a gene network model. In the resulting network architecture we identified a small subnetwork, formed by seven interconnected nodes, that controls the parasite's life cycle. The present approach could be useful for better understanding other single cell organisms with multiple developmental stages.

Citation: Carrea A, Diambra L (2016) Systems Biology Approach to Model the Life Cycle of *Trypanosoma cruzi*. PLoS ONE 11(1): e0146947. doi:10.1371/journal.pone.0146947

Editor: Luzia Helena Carvalho, Centro de Pesquisa Rene Rachou/Fundação Oswaldo Cruz (Fiocruz-Minas), BRAZIL

Received: August 6, 2015

Accepted: December 22, 2015

Published: January 11, 2016

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: The authors have no support or funding to report.

Competing Interests: The authors have declared that no competing interests exist.

Introduction

One of the main aims in the post-genome era is to elucidate the complex webs of interacting genes and proteins underlying the establishment and maintenance of cell states. Consequently, many researchers have focused on developing quantitative frameworks to identify modules that govern the transitions between different phenotypes [1–3]. The gene regulatory network (GRN) approach is one of the most popular frameworks used today [4, 5]. This approach has been used to study key reprogramming genes and cell differentiation processes in stem cells from different points of view [6–8]. Mathematically, GRN models are dynamical systems whose states determine the gene-expression levels [4]. The structure of the network is defined as a graph whose nodes are associated with genes (or groups of genes), and whose edges represent the interactions between the nodes. The task of uncovering the GRN architecture from the cell states (gene-expression profiles) represents a very complex inverse problem that has become central in functional genomics [9]. The main drawbacks of this reverse engineering task are not only the large number of genes and the limited amount of data available, but also the nonlinear dynamics of regulations, the inherent experimental errors, the noisy readouts of expression levels, and many other unobserved factors that are part of the challenge [10]. Although emerging technologies offer new prospects for monitoring mRNA concentrations, researchers have focused on determining the architecture of simplified theoretical models [11].

In this work, we have implemented a GRN approach to analyse transcriptional data of the steady states of the flagellated protozoan parasite *Trypanosoma cruzi* (*T. cruzi*). This trypanosomatid is the causative agent of Chagas disease, that affects about 7–8 million people worldwide causing about 12,000 deaths per year. Usually, the parasites are transmitted to humans and to other mammalian hosts mainly by contact with the faeces of infected blood-sucking triatomine bugs [12]. *T. cruzi* has several developmental stages both in insect vectors and in mammalian hosts (Fig 1a). Insects become infected by sucking blood from mammals with circulating parasites (trypomastigotes). In the midgut of the insect, trypomastigotes differentiate

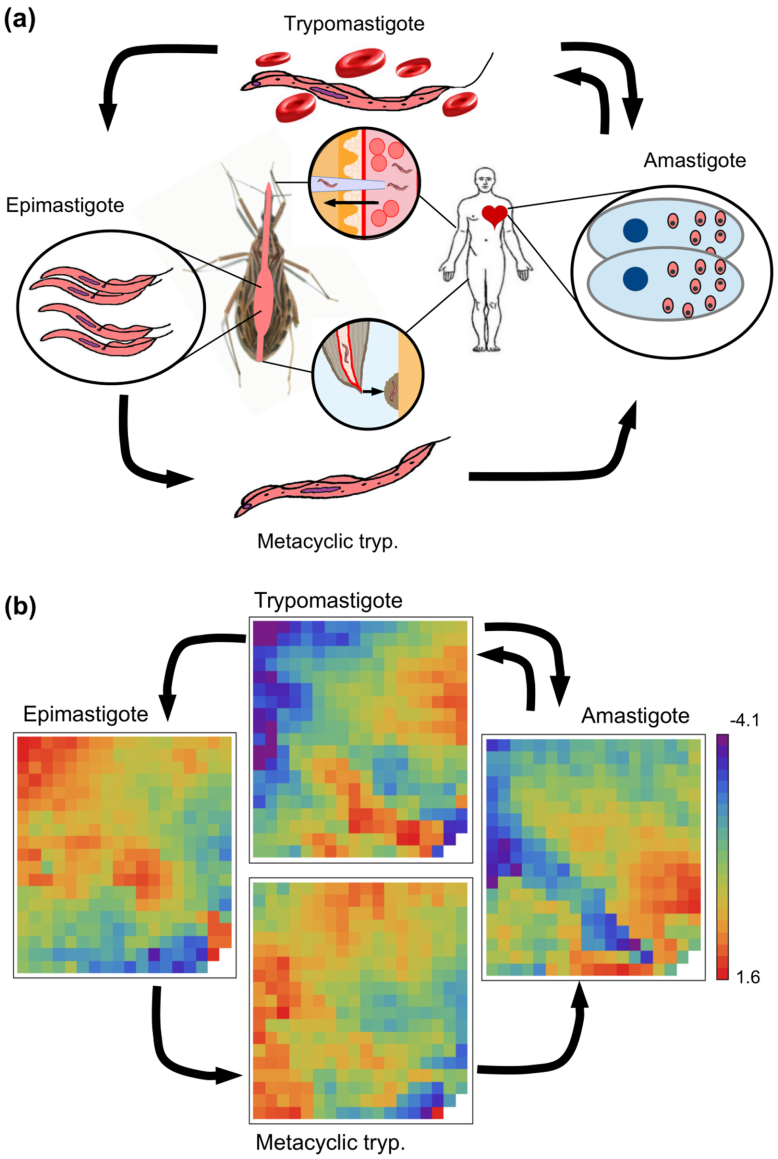

Fig 1. The life cycle of *T. cruzi*. (a) Sketch illustrating the life cycle of the parasite. (b) Plots illustrating the transcriptional snapshots of the parasite's four stages. After a dimensional reduction analysis of the microarray dataset, we have found that the four steady states can be represented by 339 variables. Each of these variables (cells in the 19 × 18 array) corresponds to the intra-cluster average of the log-transformed relative expression level of the genes that belong to the corresponding cluster. Since gene assignment to the clusters is the same for all states, the arrays can be directly compared with one another.

doi:10.1371/journal.pone.0146947.g001

into epimastigotes that replicate by binary fission. Then, epimastigotes differentiate into metacyclic trypomastigotes in the hindgut. This parasite form is released in the insects faeces and enters the mammalian host. Once in the vertebrate host, metacyclic trypomastigotes invade local cells and differentiate into amastigotes that replicate by binary fission. They subsequently transform into trypomastigotes inside the cells. By lysing the cells, trypomastigotes are released into the circulation. Thus, they spread via the bloodstream and infect new cells from distant tissues, mainly muscle and ganglion cells. There, trypomastigotes transform back into intracellular amastigotes which, in turn, undergo further cycles of intracellular multiplication. The cycle of transmission is completed when circulating trypomastigotes are taken up in blood meals by triatomine vectors [13, 14].

Even though there are potential vaccine candidates against *T. cruzi* infection, no vaccine is yet available [15]. Thus, the finding of novel therapeutic targets remains a significant challenge in the control of Chagas disease. Taking this into consideration, we have implemented a GRN approach to analyse transcriptional data of *T. cruzi*'s steady states [16]. The first analyses of networks related to *T. cruzi* were not truly GRN approaches but extracellular matrix (ECM) interactomes [17, 18]. These protein-interaction networks were built using the MiMI Cytoscape plugin. In these works some parasite surface molecules, which are known to modulate or interact with these host proteins, were pointed out.

In the present framework, we have uncovered the underlying architecture network that supports the steady states associated with the four phenotypic stages of *T. cruzi* and the transitions between the parasite's life cycle stages in response to environmental cues. We believe that this gene network model can clarify the signaling pathways, predict the response of cellular systems to multiple perturbations other than the ones used to derive the model, and determine the perturbation pattern for any desired response.

Materials and Methods

Microarray data normalisation

In this work we have used the microarray experiments of Minning *et al.* [16]. These data are publicly available in Gene Expression Omnibus (GEO) database (Accession no.: GSE14641). This series is the result of dye-swap experiments, out of which we selected the probe intensity signals of 12 microarrays (three biological replicates, and the four not-mixed parasite stages). These microarrays comprise 12,288 unique 70-mers designed against open reading frames in the annotated CL Brener reference genome sequence. They also contain 500 control oligonucleotides designed from *Arabidopsis* sequences. All of these oligonucleotides were printed in duplicate. Further details about probe preparation, microarray hybridisation, and data acquisition can be found in [16], while the description of the microarrays is available at http://pfgrc.tigr.org.

The probe intensity signals from the microarrays were subjected to the following normalisation procedure. (i) The signal intensity of each probe was set at the average of the signal intensities associated with a pair of replicate spots. (ii) The signal intensity of a probe i was normalised against the average signal of control *Arabidopsis* probes in order to obtain a signal relative intensity within the slide. The average signal of control *Arabidopsis* probes is the arithmetic mean of a set of control probes with valid signals. Of course, this set is the same in all microarray experiments. This normalisation procedure has allowed us to integrate the expression data of all the microarrays. The signal relative intensity of the probe i recorded in one of the biological replicates, $j = 1, 2, 3$, at one of the stages, $\alpha = 1, 2, 3, 4$, was represented by $y_i^{\alpha_j}$. (iii) After within-slide replicates processing, we averaged the relative intensity $y_i^{\alpha_j}$ over all replicates belonging to the same stage, i.e., $\bar{y}_i^{\alpha} = 1/3\sum_j y_i^{\alpha_j}$. Probes without a valid relative signal in all

three biological replicates were not considered in the subsequent analyses. As a result of this processing, we obtained the relative intensity of 8904 probes at parasite's four different stages. We then considered the variable that describes the expression level of the probe i at stage α as the quantity $x_i^\alpha = Log_e[\bar{y}_i^\alpha / \langle \bar{y}_i^\alpha \rangle_\alpha]$. S1 Table lists all normalised expression levels, x_i^α, used in the following analyses, with their corresponding oligo IDs.

Clustering procedure

Instead of using each gene's profile, many researchers have analysed the cell at a higher level of abstraction. One way to do this is by grouping redundant genes, i.e. by clustering co-expressed genes [19, 20], and using the average within each cluster as a variable. In order to group the genes by similar expression profiles, we have applied an agglomerative hierarchical clustering method; the Unweighted Pair Group Method with Arithmetic Mean (UPGMA), though a more sophisticated clustering method like self-organizing map could be used for this task. The agglomerative process is stopped at a given number of clusters considered suitable for our data-set. Since the suitable number of clusters, N_c, is not known, it has to be computed beforehand. In order to do this, we repeated the clustering procedure for several N_c values, and computed the Davies-Bouldin index (DBI) as a measure of the clustering merit [20, 21]. The DBI is defined as:

$$E = \frac{1}{N_c} \sum_{j=1}^{N_c} \frac{\max \left\{ \delta_k - \delta_j \right\}}{||c_k - c_j||}, \tag{1}$$

where $\delta_k = N_k^{-1} \sum_i ||x_i - c_k||$ denotes the centroid intra-cluster distances of cluster k (N_k being the number of genes belonging to cluster k); and $||c_k - c_j||$ is the distance between the cluster centroids. A low DBI value indicates a good cluster structure. It should be noted that increasing N_c without penalty will always reduce the resulting index. Then, the choice of N_c will intuitively strike a balance between the data compression and the accuracy of the dimensionality reduction. S1 Fig displays the DBI as a function of N_c for the gene-expression profile under study. It can be seen that the DBI does not suffer a significant reduction beyond N_c = 339. Thus, N_c = 339 was selected as the optimal number of clusters. The profiles of the 8,904 genes were grouped in 339 clusters, and the intra-cluster average of the expression level (i.e. $\bar{x}_j^\alpha = \langle x_i^\alpha \rangle_{i \in \text{cluster} j}$) was used in all of the subsequent analyses. Fig 1b displays a 2D array of the resulting average levels after the dimension reduction process described above for the four stages of *T. cruzi*'s life cycle. The sets of genes belonging to each cluster are listed in S2 Table, and the intra-cluster averages of the expression levels, $\bar{x}_j^\alpha$, for each cluster (rows) at each of the parasite's stages (columns) are listed in S3 Table.

Reverse engineering methods

Gene network dynamics. In this work, we have implemented a discrete-time linear model [22–24] which has two advantages: it can take into account fluctuations, and its parameter estimation does not involve intensive computational steps [11]. In this model, the system's state at time t is represented by an N-dimensional vector $\mathbf{x}(t)$, which represents the activity of the N nodes of the network. The temporal evolution of the gene network is governed by:

$$x_i(t + \Delta t) = \sum_j w_{i,j} x_j(t) + \theta_i + k_i^\mu + \epsilon_i(t), \tag{2}$$

where $w_{i,j}$ are the elements of the weighted connectivity matrix $\mathbf{W}$, θ_i is a constant bias term of gene i, and k_i^μ determines the influence of the environmental cue μ on gene i. We have

considered four different cues corresponding to unknown external differentiation signals. Thus, $\mu = 1, 2, 3$, and 4 represent the external signals responsible for the transitions to the amastigote, epimastigote, metacyclic tryp., and trypomastigote stages, respectively. Finally, $\epsilon_i(t)$ is a noise term assumed to be Gaussian with mean equal to 0.

In order to simplify the notation for the parameter estimation procedure, we noticed that the bias term and the environmental cues can be included in an extended version of matrix **W** and of state vector **x**. Thus, the state of gene *i* is given by:

$$x_i(t + \Delta t) = (w_{i,1}, w_{i,2}, \ldots, w_{i,N}, \theta_i, k_i^{\mu}) \cdot (x_1, x_2, \ldots, x_N, 1, 1) + \epsilon_i, \tag{3}$$

where μ corresponds to the acting environmental cue. This said, the same parameter estimation method can be applied whether the environmental cues are present or not.

Singular value decomposition (SVD). Linear models serve as the basis of all continuous gene-network approaches currently available to model typical time-course gene-expression data sets (see [11] for a review). These data sets consist of M pairs of input-output states, represented by $D = \{\mathbf{X},\mathbf{Y}\}$. Matrix **X** is the $N \times M$ gene-expression matrix at time t. The columns of matrix **X** labeled by index v, $\mathbf{x}^v$, correspond to the experiments, while the rows indicate individual genes. The same is valid for the gene-expression matrix at time $t+\Delta t$, **Y**. For a given D, the linear model must map each gene-expression state to the consecutive state, i.e.:

$$\mathbf{y}^v = \mathbf{W}\mathbf{x}^v, \qquad v = 1, \ldots, M. \tag{4}$$

Therefore, in order to find the connectivity matrix, the predicted states from a given input state $\mathbf{x}^v$ of the training set must be as close as possible to the output state $\mathbf{y}^v$. An alternative would be to minimise the cost function $\Sigma_v \|\mathbf{W}\mathbf{x}^v - \mathbf{y}^v\|$. A particular solution with the smallest L_2 norm is given in terms of the SVD of matrix $\mathbf{X}^{\mathrm{T}}$ (where superscript T denotes the transpose matrix), i.e. $\mathbf{X}^T = \mathbf{U} \cdot \mathbf{S} \cdot \mathbf{V}^T$, where **U** is a unitary $M \times N$ matrix of left eigenvectors, **S** is a diagonal $N \times N$ matrix containing the eigenvalues $\{s_1, \ldots, s_N\}$, and **V** is a unitary $N \times N$ matrix of right eigenvectors [23, 25]. Thus, the solution with the smallest L_2 norm represented by $\mathbf{W}_{L_2}$ is given by:

$$\mathbf{W}_{L_2} = \mathbf{Y} \cdot \mathbf{U} \cdot \mathrm{diag}(s_j^{-1}) \cdot \mathbf{V}^T. \tag{5}$$

Without loss of generality, all s_j elements whose value is different from 0 were listed at the end of diagonal matrix **S**, and the s_j^{-1} values in Eq (5) were considered to be 0 if $s_j = 0$.

SVD is mathematically related to the eigen decomposition (ED) and the principal component analysis (PCA). SVD can be understood as a generalisation of ED, but ED only applies on diagonalizable matrices, and fails if the matrix is not square or is singular as in our case. SVD and ED are related, in particular the column vectors of **U** are eigenvectors of $\mathbf{X} \cdot \mathbf{X}^T$, while the column vectors of **V** are eigenvectors of $\mathbf{X}^T \cdot \mathbf{X}$. The diagonal entries of **S** are the square roots of the eigenvalues of both $\mathbf{X} \cdot \mathbf{X}^T$ and $\mathbf{X}^T \cdot \mathbf{X}$. The eigenvectors of $\mathbf{X}^T \cdot \mathbf{X}$ also allow to compute the principal components in the standard PCA. PCA is a common tool for exploratory data analysis. It can provide a lower-dimensional picture by projecting the data onto the subspace with the higher variance [26]. We have exploited PCA analysis for a better visualisation of our results. However, since the focus of this manuscript is finding the connectivity matrix rather than obtaining a better representation of data sets, we have used SVD instead of ED or PCA because it is more suitable for that purpose.

The smallest L_2 norm solution cannot be unique. Assuming that $\mathbf{x}^v$ are linearly independent, finding the unique solution requires that $M \geq N$. Unfortunately, the inverse problem in GRN involves $M << N$. Thus, the problem tends to be severely underdetermined, and many solutions can then be consistent with data D. Therefore, all the possible connectivity matrices that

are consistent with Eq (4) can be written in a closed form as:

$$\mathbf{W} = \mathbf{W}_{L_2} + \mathbf{C} \cdot \mathbf{V}^T, \tag{6}$$

where $\mathbf{C}$ is an $N \times N$ matrix whose elements c_{ij} are 0 as long as $s_j \neq 0$. Otherwise, they are arbitrary scalar coefficients. As it will be seen later, the degrees of freedom due to this arbitrariness can be exploited to our benefit [23]. The solution offered by Eq (5) is implemented to embed the four steady states of *T. cruzi* into the dynamics of the model, without considering the transitions between the states. Eq (6) is used to uncover the environmental cues by means of using the information provided by the transitions between the different stages of the parasite's life cycle, and the connectivity matrix associated with the steady states inferred in the previous step.

Embedding the steady states. In order to infer the parameter values of Eq (4) that allow the model to display the same set of steady states as the ones seen in the parasite, we have constructed a training set of size M, represented by D_{ss}. Different noise realisations associated with the four stages were added to the steady states. Thus, the columns of matrix $\mathbf{X}$ are given by:

$$\mathbf{x}^v = \{\bar{x}_j^\alpha\} + \{\epsilon_j^i\}, \quad \text{with } j = 1, \ldots, N \text{ and } v = 1, \ldots, M,$$

where $\alpha = 1, 2, 3, 4$, and the superscript i denotes the noise realisations. In this work, we have used 40 noise realisations for each steady state. Thus, $M = 4 \times 40 = 160$. ϵ_j is a Gaussian noise with mean equal to 0 and a small standard deviation (set at 1% of the expression data). The columns of matrix $\mathbf{Y}$ ($\mathbf{y}^v = \{\bar{x}_j^\alpha\} + \{\epsilon_j^{i'}\}$) are defined in the same way. We have expanded the size of the training set, thereby making the solutions more robust against the fluctuations. This simple concept is similar to adding a Tikhonov regularisation term in the optimisation process, which has been studied in several neural network problems [27, 28]. The constructed training set implies that if at a given time the system is very close to one steady state, it will remain close to that steady state in the next time-step as well (Fig 2c).

Overfitting occurs when the model has a good performance on the training set, but it has a poor generalization performance, i.e. a poor ability to correctly predict data beyond the training set. There are several reasons for a model to adjust very specific random features of the training set, with a consequently poor generalisation power [29]. Among them, it can be mentioned models with many parameters and small training set, or when trying to learn training examples that have no causal relationship to the target function. The latter situation is more common in regression problems such as the one considered here. For example, overfitting was reported using SVD when recovering a gene network from a highly noisy training set (20% of the expression level), but it was not present when recovering the network from a clean training set [23]. The noise level used to generate training set D in this work is very low (<1%), and does not cause overfitting.

In order to discriminate if an estimated matrix element should be 0 or another reliable value different from 0, we have constructed not only a training set, D_{ss}, but an ensemble of training sets by means of using different noise realisations. For each training set we have computed the minimal L_2-norm solution. Different noise realisations give slightly different solutions. Thus, the ensemble of solutions defines a probability distribution for each weight, $P_{i,j}(w)$. We then performed a location test for each distribution $P_{i,j}(w)$, as illustrated in Fig 2d. This step consists of testing the hypothesis stating that the true mean value of $P_{i,j}(w)$ differs from 0 at some magnitude (set at 0.0075). If the p-value associated with this test is greater than 0.01, the hypothesis is rejected, and $w_{i,j}$ is assigned 0 value. Otherwise, $w_{i,j}$ is assigned the mean value of $P_{i,j}(w)$. This procedure allows us to obtain a sparse connectivity matrix, $\mathbf{W_{ss}}$, that is compatible with the steady states.

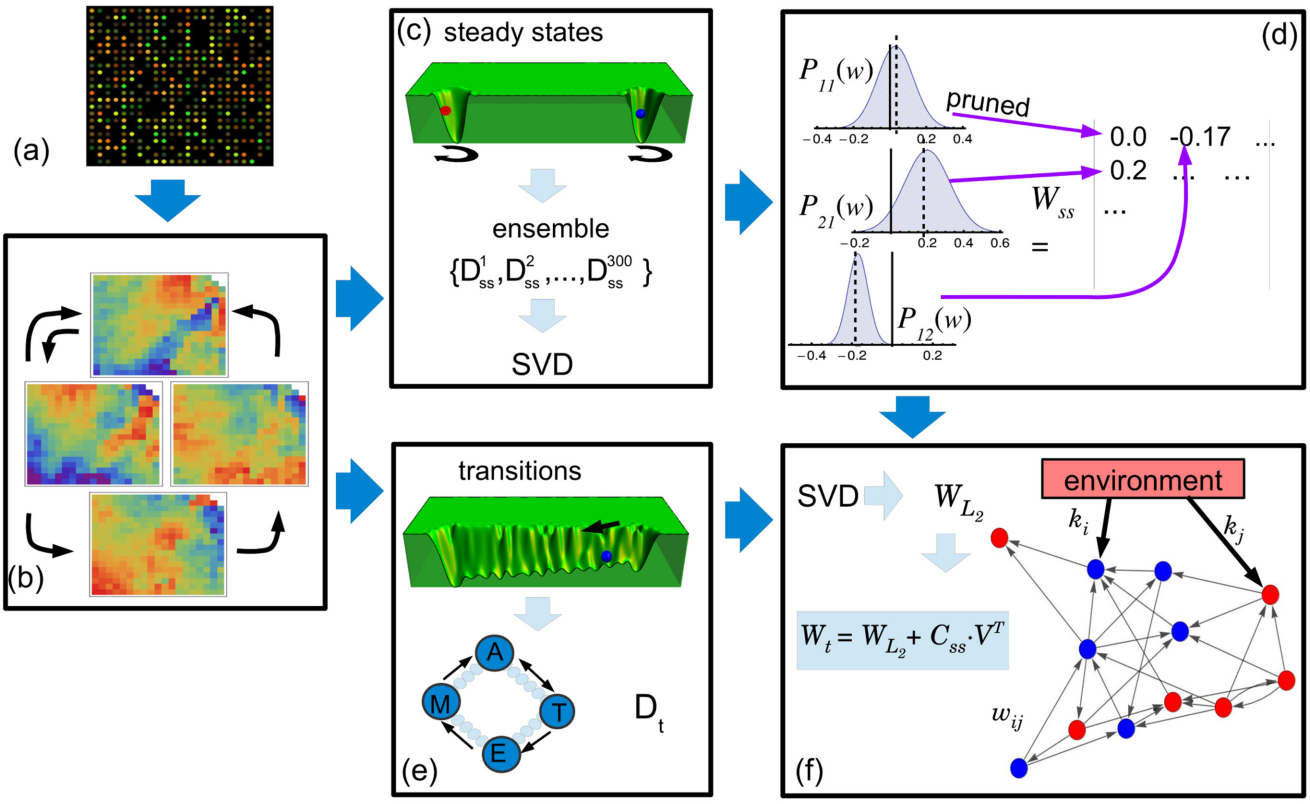

Fig 2. Schema of the network inferring method. (a) The microarray data corresponding to the parasite's four steady states are normalised. (b) The total of 8,904 gene-expression levels of each stage is reduced to 339 clusters representing the variables of our systems. (c) An ensemble of 300 training sets including fluctuations around the steady states is constructed from the steady states. Using singular value decomposition (SVD), the minimal L_2-norm solution for each D_{ss} is determined. (d) A sparse connectivity matrix, $\mathbf{W_{ss}}$, is derived from the probability distribution $P_{ij}(w)$ by using a pruning method based on a location test. (e) A new training set is constructed from the transitions between the amastigote (A), epimastigote (E), metacyclic tryp. (M) and trypomastigote (T) stages. Intermediate states (small circles) between the stages are assumed to exist. It is also considered that an unknown external cue (black arrow) is responsible for the transitions. (f) By means of using SVD, the L_2-norm solution, $\mathbf{W_{L_2}}$, is determined. This solution is in turn used to find another solution, $\mathbf{W_t}$, which includes information concerning the steady states. This procedure is used to infer the weighted links between genes, $w_{i,j}$, and to answer two questions: which genes are affected by the external cues, and how they are regulated (up or down) by the environment.

doi:10.1371/journal.pone.0146947.g002

Embedding the transitions between the steady states. In order to extend our analysis by including the environmental cues, we have used the extended versions of **W** and **x** described by Eq (3). To embed the transitions between the steady states, we have considered that these transitions occur gradually and through the shortest possible path between the steady states. Thus, if the system is in the steady state $\mathbf{x}^{\alpha}$ and is driven to the steady state $\mathbf{x}^{\beta}$ due to an external cue $\mu = \beta$, then the system performs a series of small transitions between intermediate states represented by $\mathbf{x}^{\alpha,\beta}(t)$. These intermediate states were constructed by means of a linear combination of the initial and final steady states, i.e. $\mathbf{x}^{\alpha,\beta}(t) = ((n_i - t)\ \mathbf{x}^{\alpha} + t\ \mathbf{x}^{\beta})/n_i$ with $t = 0, 1, 2, \ldots, n_i$. As it can be seen, $\mathbf{x}^{\alpha,\beta}(0)$ and $\mathbf{x}^{\alpha,\beta}(n_i)$ coincide with the steady states $\mathbf{x}^{\alpha}$ and $\mathbf{x}^{\beta}$, respectively. Thus, using these intermediate states, we constructed a new training set D_t, where the columns of matrices **X** and **Y** are defined as follows:

$$\mathbf{x}^{\nu} = \{\bar{x}_j^{\alpha,\beta}(t)\} + \{\epsilon_j^i\}, \quad \mathbf{y}^{\nu} = \{\bar{x}_j^{\alpha,\beta}(t+1)\} + \{\epsilon_j^{i'}\},$$

where $t = 0, 1, 2, \ldots, n_i - 1$. We have used $n_i = 10$, which implies 10 small transitions. The pairs (α,β) correspond to the allowed transitions between the steady states; five transitions in the case of *T. cruzi*. Again, ϵ_j is a Gaussian noise with mean equal to 0 and a small standard

deviation (set at 1% of the expression data). In all, four different noise realisations were used, and the size of our training set, D_t, was in turn $M = 200$. We then computed its smallest L_2 norm solution, $\mathbf{W}_{L_2}$. However, since $M < N$, this solution was not unique. In order to find a particular solution as close as possible to connectivity matrix $\mathbf{W}_{ss}$, we used Eq (6) and computed the elements of matrix $\mathbf{C}_{ss}$ that obey the following equation:

$$\mathbf{W}_{ss} = \mathbf{W}_{L_2} + \mathbf{C}_{ss} \cdot \mathbf{V}^T, \tag{7}$$

where matrix $\mathbf{W}_{ss}$ was padded with 0 values because $\mathbf{W}_{L_2}$ includes four additional rows and columns corresponding to the environmental cues that are not present in $\mathbf{W}_{ss}$. Eq (7) is an overdetermined problem that can be solved by applying the interior point method for L_1 regression [23]. The resulting c-values were then used to compute a new connectivity matrix (Fig 2f). This matrix, represented by $\mathbf{W}_t$, is not only consistent with the information of the environmental cues and transitions included in D_t, but it is also close to $\mathbf{W}_{ss}$.

Results

GRN modeling

Key decisions in modeling a gene network system include the choice of variables and the mathematical framework for representing the system dynamics. In this sense, several regulatory network approaches such as Bayesian networks [30], Boolean networks [31], and linear models [22, 24, 32] have been suggested. The model must be chosen based on the available data and the ability to infer accurate-enough parameters. The more detailed the model, the more experimental data required to make it work. For instance, when choosing a linear model, in which the expression levels of N genes at time t determine the changes of such expression levels at time $t + \Delta t$, the transition matrix must be computed from N pairs of input-output data.

In this work, we have assumed that the system's state is represented by $\mathbf{x}(t)$ –the N-dimensional vector corresponding to the expression levels of N gene clusters measured at time t. The GRN dynamics is modeled by a first order Markov model, where the future state depends linearly on the present state and on external perturbations. Mathematically, it is defined by the following equation:

$$x_i(t + \Delta t) = \sum_j w_{i,j} x_j(t) + \theta_i + k_i^{\mu} + \epsilon_i(t), \tag{8}$$

where $w_{i,j}$ are the elements of the weighted connectivity matrix $\mathbf{W}$, and indicate the type and strength of the influence of gene j on gene i ($w_{ij} > 0$ indicates activation, $w_{ij} < 0$ indicates repression, and 0 indicates no influence). θ_i is a constant bias term to capture the activity level of gene i in the absence of regulatory inputs. We have also added a term indicating the influence of unknown external perturbations, or environmental cues; k_i^{μ}, which is the influence of the environmental cue μ on gene i. Finally, $\epsilon_i(t)$ is a noise term assumed to be Gaussian with mean 0. The next task in our work was to determine which nodes were affected by external cues –even if those cues were unknown–, and how they were affected. To this end, we considered not only the expression-profile data set information ($\bar{x}_j^{\alpha}$), but also some *a priori* information associated with the following biological facts: (i) the parasite's life cycle has four stages, each of them associated with a measured steady state; (ii) each steady state exhibits some level of noise or fluctuations; and (iii) there are five possible transitions between these four stages. We have assumed that these transitions are the result of different environmental cues acting on certain nodes of the network. Following these facts, we implemented a two-step reverse engineering protocol sketched in Fig 2. First, we focused on embedding the four steady states into the dynamics of the

model, regardless of the transitions between these states. Second, we concentrated on uncovering the environmental-cue effectors considering the transitions between the parasite's life cycle stages, while using the same connectivity matrix derived in the previous step.

Modeling the steady states of *T. cruzi*

In order to infer connectivity matrix **W**, we have considered the linear model (Eq (8)) without external perturbations, and have applied the SVD procedure over a training set, D_{ss}, constructed as indicated in Methods. As a result, the dynamical system, together with the derived matrix, has four basins of attraction which correspond to each of the parasite's stages. This means that whenever the system is in a given basin, it will remain inside that basin as long as there are no external perturbations.

One way to quantify how close the system is of a particular state is by computing the overlap between the vectors that represent the actual and the target stage, where the overlap between vectors x and y is mathematically defined as $x \cdot y/|x||y|$. Fig 3a depicts the trajectories (black lines) that illustrate the dynamics of our model in the space spanned by the three principal components. This plot shows that the trajectories fluctuate around corresponding stages represented by colored circles (amastigote in blue, epimastigote in red, metacyclic tryp. in green, and trypomastigote in yellow). For each trajectory we have computed the overlap between the vectors corresponding to the state of the system at time t, and the ones corresponding to each target stage. Fig 3b depicts the time course of such overlaps illustrating, in a more quantitative manner, how the trajectories fluctuate around the the corresponding stages. Fig 3c shows a 2D schematic illustration of the pseudo-potential landscape with the four basins of attraction.

The elements of matrix **W** are continuous variables and, consequently, they are associated with not-null values. However, the statistical analysis of known regulatory networks has

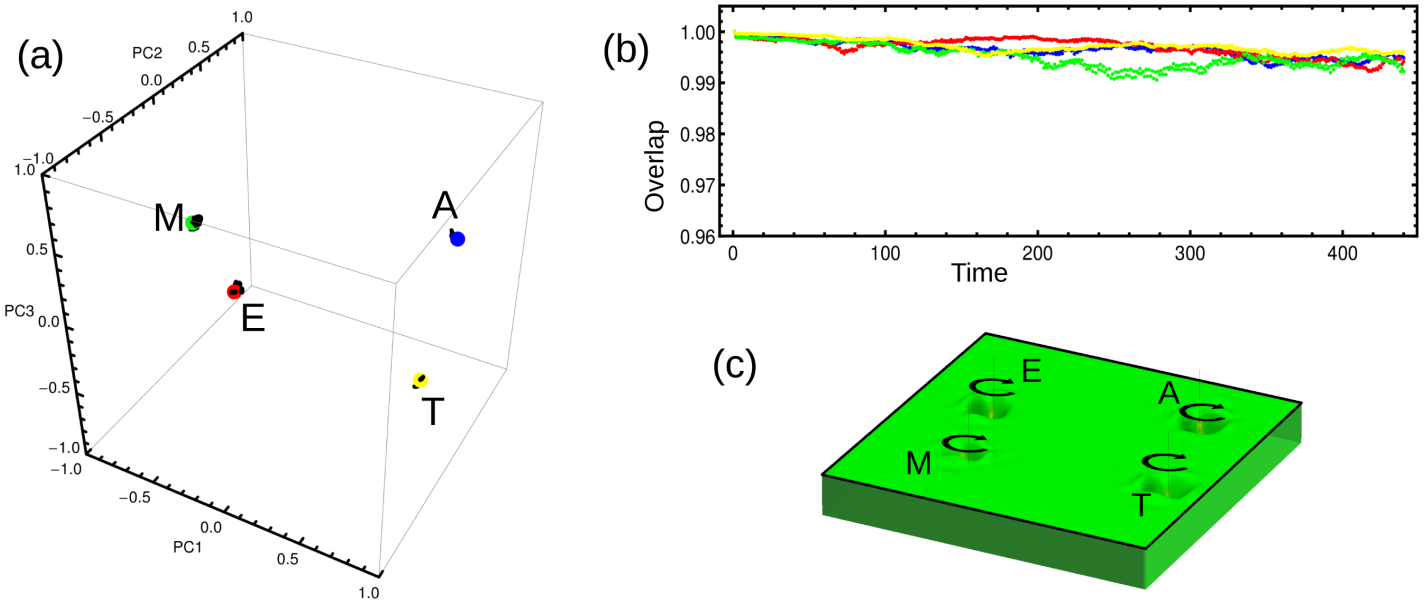

Fig 3. Stability of the steady states. (a) The plot shows the positions of the four steady states of the parasite's life cycle in the space spanned by the three principal components. The black trajectories around each stage are the result of simulations conducted using the model (Eq (8)) without external cues. A slightly perturbed steady state was used as the initial condition. The system fluctuates around the corresponding steady state. (b) Temporal behavior of the overlap between the state of the system at time t and the amastigote steady state (blue), the epimastigote steady state (red), the metacyclic tryp. steady state (green), or the trypomastigote steady state (yellow). (c) 2D projection of the pseudo-potential landscape with the four basins of attraction corresponding to each of the parasite's stages. The circular black arrows represent the system's fluctuation around the steady states, just as seen in Fig 3a.

doi:10.1371/journal.pone.0146947.g003

revealed that such networks have a sparse nature, i.e. the number of actual edges in a network is very small compared to the number of possible edges [33, 34]. Such sparsity is difficult to obtain when dealing with continuous weights. Thus, the inferred matrix elements at the end of the reverse engineering process should be either 0 or another reliable value different from 0.

In the spirit of inferring a sparse weight matrix that allows the system to display the four steady states, we have consider a bootstrap method. In this sense, an ensemble of 300 training sets was constructed by means of adding different noise realisations to the steady states, as described in Methods. Using SVD we computed a solution for each training set, obtaining a probability distribution for each weight, $P_{i,j}(w)$. The next step was to assign a value to each element of the connectivity matrix, while carefully assessing the significance of the weight values. We performed a location test to prune the non-significant weights, as illustrated in Fig 2d, and constructed a sparse connectivity matrix, $\mathbf{W}_{ss}$, which supports the data set. At the significance level of 0.01 there are 11,470 links between genes, i.e. around 10% of the elements of $\mathbf{W}_{ss}$ are not null. Even with this average node degree, the visualisation of the resulting network poses a challenge. In order to overcome this difficulty, we have displayed only a small fraction of the nodes (around 470 links with p-value less than or equal to 10^{-200}). Fig 4 shows the GRN. The

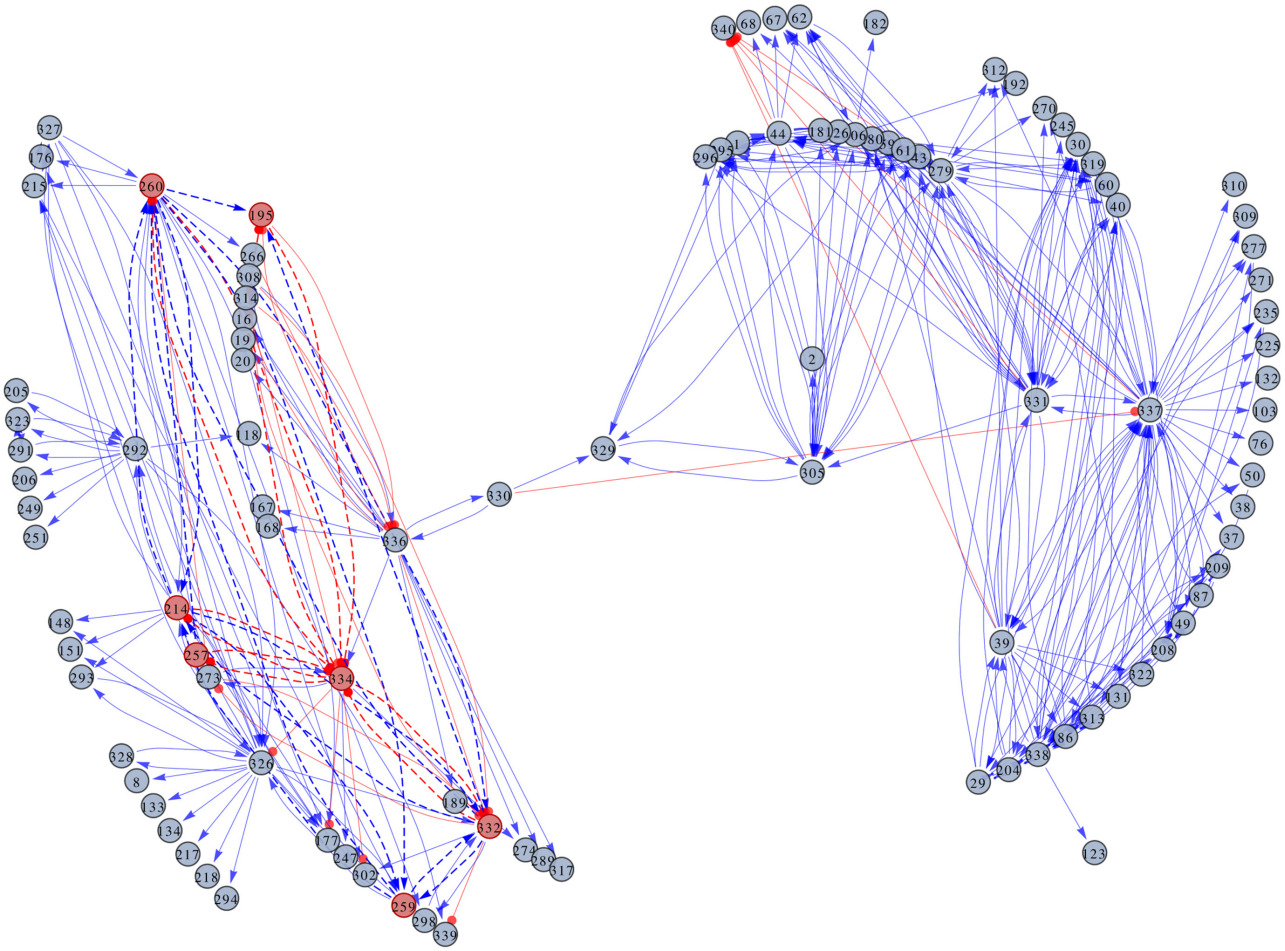

Fig 4. GRN representation of the steady states of *T. cruzi*. The network edges represent the regulatory links between the gene clusters, while the nodes represent the clusters themselves. The labels inside the nodes correspond to the cluster IDs. Additional information about the clusters can be found in S5 and S7 Tables. The regulatory links indicate either the activation (arrows) or the repression (lines ending in circles) of the clusters. A seven-node subnetwork that controls the dynamics of the parasite's life cycle is highlighted.

doi:10.1371/journal.pone.0146947.g004

two weakly connected subnetworks seen in the graph reveal a modular organisation of the network at the significance level used. As it will be seen later, one of these subnetworks is linked to the parasite's life cycle. The 11,470 links (all not-null elements of the matrix) between genes are listed in S4 Table.

Extracting valuable information from a network made of 10,000 links is a complex task. One way to overcome this problem is by considering only the more important regulators of each steady state. Since the whole regulatory output of a gene depends on the gene's activity level, some genes can be important regulators in one state, while their activity level in the other three states is low (i.e. $x_i \sim 0$). With this in mind, we have constructed network plots that emphasise the most important links in each steady state; that is to say, those links with $|w_{i,j} x_j| \geq 5\%$ of $|x_i|$. The plots in S2 Fig. depict the link-derived networks for each of the parasite's four stages: amastigote (S2a Fig), epimastigote (S2b Fig), metacyclic tryp. (S2c Fig), and trypomastigote (S2d Fig). As it can be seen, some clusters present regulatory activity only in one particular state. For example, clusters 302, 308 and 333 only appear as relevant regulators in the metacyclic tryp. state, the epimastigote state and the trypomastigote state, respectively. Other clusters, however, are important regulators in all four steady states, as is the case of clusters 326, 336 and 337. Detailed biological information about the genes belonging to the more relevant clusters is listed in S5 Table. After analysing the data obtained for each of *T. cruzi*'s four stages, we have found 47 clusters with important regulatory activity. These clusters include a total of 68 genes: 25 encoding uncharacterised proteins, and 43 coding for proteins with known functions. Among the latter, the most abundant proteins are *trans*-sialidase (TS) (encoded by nine different genes), amastin (encoded by five different genes), and mucin TcMUCII (encoded by four different genes).

Besides the main four basins of attraction linked to the known steady states displayed in Fig 3, the system dynamics might include other basins of attraction not-linked to known phenotypes. An exhaustive search for these spurious attractors was performed, and another 20 small attractors, where the system can be trapped were found. These attractors have small basins associated, that disappeared when the phenotypic transitions due to the external cues are included in the model.

Modeling the phenotypic transitions of *T. cruzi*

After embedding the steady states of the parasite into the GRN dynamics, our analysis was extended to include the transitions that take place between those states as a result of environmental cues. The fact that these transitions in the presence of a given external perturbation occur gradually was taken into account. Since no data about the intermediate states between the steady states are available, we have constructed a training set, represented by D_t, considering that the system performs transitions between the initial and final steady states through the shortest possible path. For details about the construction of the training set, see Methods. As this training set has $M < N$, there exist infinite solutions compatible with D_t. We have chosen the closest solution to the connectivity matrix that uses nothing but the steady states information, i.e. the closest to $\mathbf{W}_{ss}$. Thus, our connectivity matrix is represented by:

$$\mathbf{W}_t = \mathbf{W}_{L_2} + \mathbf{C}_{ss} \cdot \mathbf{V}^T, \tag{9}$$

where $\mathbf{W}_{L_2}$ is the corresponding minimal L_2-norm solution obtained by SVD for D_t. Matrix $\mathbf{C}_{ss}$ was computed by the interior point method as described in Methods. The new connectivity matrix, $\mathbf{W}_t$, is consistent with the information of the environmental cues and transitions included in training set D_t. As $\mathbf{W}_t$ is also very close to $\mathbf{W}_{ss}$, it consequently inherits the ability to support the multi-stability of the parasite's life cycle.

In order to test the ability of the model (Eq (8)) to emulate the observed dynamical behavior, simulations under different external cues were performed. Each of these simulations was performed considering that the system is initially in one of the parasite's steady states, and that an external cue μ is acting. The simulations were performed by running 12 iterations of the model (Eq (8)), and recording the system's state at each of these 12 steps. The temporal evolution of the 339 variables of the system was compiled in movies available as S1–S5 Movies. S1 Movie shows the simulated phenotypic transition from the amastigote stage to the trypomastigote stage when external cue $\mu = 4$ is acting. S2–S5 Movies, on their part, illustrate the modeling results of the remaining phenotypic transitions. In all cases, the final state of the system is in agreement with the expected state regarding the acting external cue. This agreement can be better appreciated when using a principal component analysis procedure for dimension reduction. Fig 5 depicts a set of trajectories corresponding to four of the five phenotypic transitions in the 3D space spanned by the main principal components. There are 20 alternative trajectories for each simulated phenotypic transition. All of the trajectories for a given transition have the same initial condition, are affected by the same external cue, but present particular noise realisations. Hence, it can be said that the model is able to reproduce the dynamics of *T. cruzi*'s life cycle.

In our model, the phenotypic transitions are caused by an environmental cue μ through parameter k_i^μ, i.e. the gene clusters associated with large positive (or negative) k_i^μ values are activated (or inhibited) by the acting external cue μ. $\mathbf{k}^\mu$ values are distributed around 0. In order to identify the key connections that modulate the network behavior under external cues, we have selected those gene clusters with $\mathbf{k}^\mu$ values greater (lower) than the 95th (5th) percentile of the distribution. These clusters are listed in S6 Table. A total of 166 externally regulated genes belonging to 86 different clusters were found. While 73 of these genes encode uncharacterised proteins, the other 93 genes code for proteins with known functions. Just as in the steady states, the most abundant proteins are TS (encoded by 21 different genes), amastin (encoded by six different genes), and mucin TcMUCII (encoded by five different genes). The difference, however, is that when considering the transitions, these proteins act no longer as regulators, but

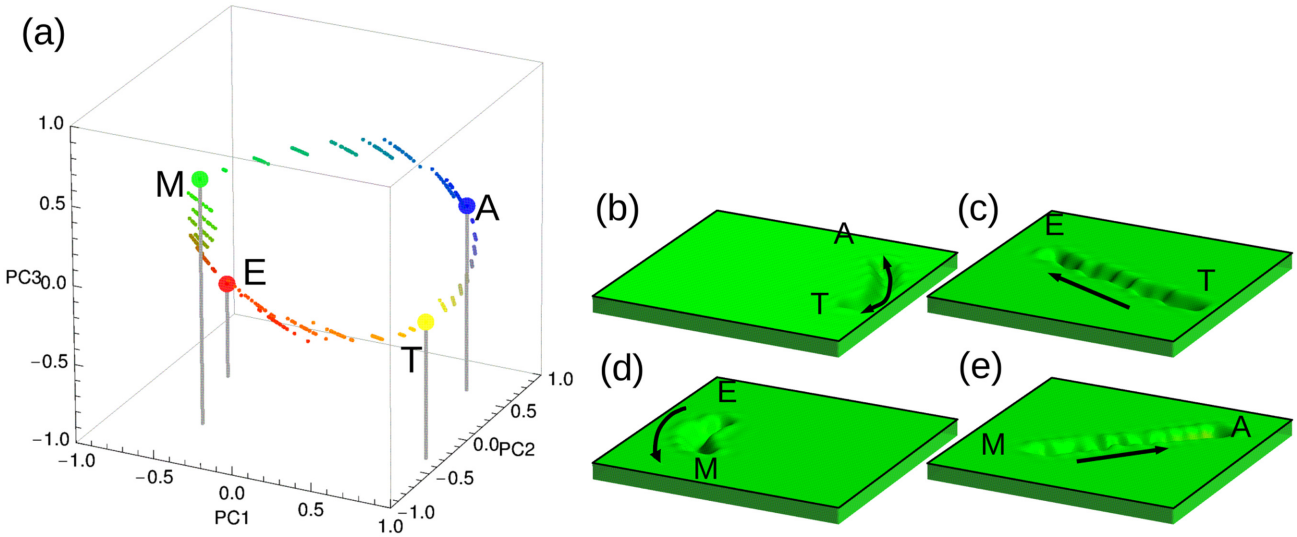

Fig 5. Representation of transitions between the steady states caused by external cues. (a) The plot shows the trajectories of the system from an initial to a final steady state under the influence of an external cue in the space spanned by the three principal components. A slightly perturbed steady state was used as the initial condition. Since amastigote-to-trypomastigote and trypomastigote-to-amastigote transitions overlap, only the first one is shown. Each trajectory has 10 intermediate states represented by small circles. (b), (c), (d) and (e) 2D projections of the pseudo-potential landscapes corresponding to the phenotypic transitions mentioned above.

doi:10.1371/journal.pone.0146947.g005

they are up- or down-regulated instead. Uncharacterised proteins without GO annotations have been analysed using the InterproScan software [35], and the results are summarised in S7 Table. According to this analysis, 39% of these proteins are membrane proteins. Furthermore, we have found that genes affected by the external cues leading to the parasite's two mammalian stages are inversely regulated, i.e. if they are up-regulated in one of these two stages, so they are down-regulated in the other, and vice versa. Regarding genes affected by the external cues leading to the parasite's two insect stages, we have found that they are equally regulated, i.e. they are up-regulated in both stages or down-regulated in both stages.

The next step in our analysis was to identify the module that controls the parasite's life cycle. This is a difficult task because it involves the isolation of a small subset of nodes and regulatory connections out of a network of 10,000 links. In principle, the number of possible subnetworks within a network of such a size is very large. Consequently, the evaluation of the subnetworks' dynamics is not possible. In order to solve this problem, we reduced the search space. With this in mind, we considered only those circuits that involve nodes with important regulatory roles. To this end, we have used the list of regulatory clusters shown in S5 Table, and have written a script to search for cyclic graphs, i.e. closed loops, containing such nodes in matrix $\mathbf{W_t}$. With this set of modules, we then searched for those subnetworks with the ability to emulate the parasite's dynamics. At this point, our model had to be simplified. In order to evaluate the dynamics of the system, we have considered that variable x_i is a Boolean variable, and that the system's evolution is given by:

$$x_i(t+\Delta t) = \mathrm{Sign}\left(\sum_{j'} w_{i,j} x_j(t) + \theta_i + k_i^{\mu}\right), \tag{10}$$

where index j' in the sum runs only over the nodes belonging to the module under evaluation. The parameter values are taken from $\mathbf{W_t}$, and listed in S8 Table.

As a result of the searching process, we were able to identify a seven-node module, containing a total of nine genes. Fig 6a illustrates the architecture of this subnetwork. The seven clusters forming the subnetwork are: 195, 214, 257, 259, 260, 332, and 334. Relevant information about these clusters and their composing genes is shown in Table 1. Three of the nine genes code for uncharacterised proteins (Q4DTV8, Q4DVU8 and Q4E589). According to their GO annotations, Q4DTV8 has hydrolase activity (acting on carbon-nitrogen -but not peptide-bonds, in linear amidines), Q4DVU8 has transporter activity, and Q4E589 has catalytic activity. The other six genes code for: an hexokinase (Q4D3P5), a δ-1-pyrroline-5-carboxylate dehydrogenase (Q4DRT8), a quinone oxidoreductase (Q4DHH8), a glutamate dehydrogenase (Q4DWV8), a peptidyl-prolyl cis-trans isomerase (Q4E4L9), and a metaciclina II (Q4E2M3).

The identified subnetwork reproduces many important dynamical features observed in the life cycle of *T. cruzi*. On the one hand, the phenotypic transitions from epimastigote to metacyclic tryp., from amastigote to trypomastigote, and from trypomastigote to epimastigote are reproduced under the influence of the corresponding external cue. And on the other, the phenotypic stages epimastigote, metacyclic tryp., and trypomastigote correspond to steady states of the subnetwork's dynamics. As an example of the subnetwork's dynamics, Fig 6b illustrates the basin of attraction of the module under the action of environmental cue $\mu = 4$. This external signal leads the network to the trypomastigote state. The figure shows that regardless of the initial state (there are 128 Boolean states), the final stop of the trajectories in the Boolean space is always the trypomastigote stage. Similarly, when the environmental cue is $\mu = 2$ or $\mu = 3$, the obtained basin of attraction is the epimastigote or the metacyclic trypomastigote stage, respectively.

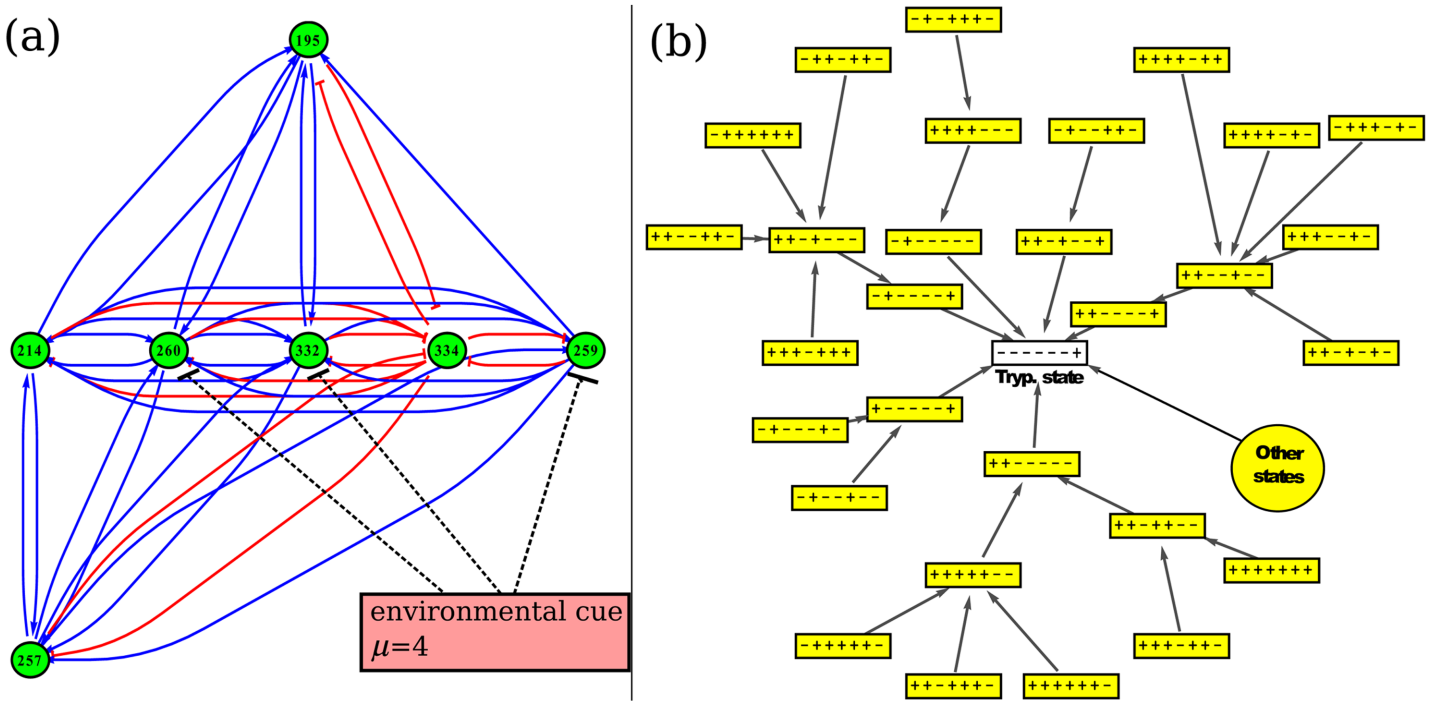

Fig 6. Life cycle module. (a) Architecture of the seven-node subnetwork linked to the parasite's life cycle. The action of environmental cue $\mu = 4$ is shown as an example. (b) Boolean dynamics of the life cycle module. The basin of attraction of the seven-node module under the action of environmental cue $\mu = 4$ is shown. This external signal leads the network to the trypomastigote state. Here, the nodes represent the module states and the edges represent the transitions. The module states are characterised by the sign of the clusters, which in turn are arranged in box according to their cluster IDs. Under the action of this perturbation, the final state is always the trypomastigote stage (white box). Some states reach this final state by going through different intermediate steps, while others (represented by the biggest circle) reach it in only one step.

doi:10.1371/journal.pone.0146947.g006

Finally, we performed two perturbation experiments in our modelling, in order to do an *in silico* validation. In this sense, we first considered the case when genes belonging to cluster 326 were overexpressed and genes belonging to clusters 337 were knocked down. These genes were predicted to have major roles for the trypomastigote stage maintenance (see S2 Fig). S3 Fig illustrates the dynamics of the system (yellow trajectory) under this perturbation. It can be seen that the trypomastigote stage is not associated to a stable attractor anymore. Instead, the system

Table 1. Subnetwork information.

Cluster ID	Gene ID	Uniprot ID	Putative function	μ	k^{μ} coefficient
195	6382	Q4DTV8	hydrolase activity	3	-0.1805
195	6588	Q4D3P5	hexokinase	3	-0.1805
214	121	Q4DRT8	delta-1-pyrroline-5-carboxylate dehydrogenase	2	0.4251
257	519	Q4DHH8	quinone oxidoreductase	2	0.2256
259	8762	Q4DVU8	transporter activity	2	0.3033
260	4053	Q4E589	catalytic activity	2	0.3479
260	5564	Q4DWV8	glutamate dehydrogenase	2	0.3479
332	4595	Q4E4L9	peptidyl-prolyl *cis-trans* isomerase	3	-0.304
334	1518	Q4E2M3	metaciclina II	2	-0.2685

List of the clusters and genes forming the parasite's life cycle subnetwork, the protein function, the most relevant external cue, and the corresponding values of $\mathbf{k}^{\mu}$ coefficients needed to reproduce the dynamical features of the system.

doi:10.1371/journal.pone.0146947.t001

moves to an alternative stable state. In the second *in silico* experiment, we simulated the effect of overexpressing genes belonging to clusters 259, 260, and 332; which were predicted to be down-regulated by the external cue leading to the trypomastigote stage (see Fig 6a). The blue trajectory in S3 Fig does not ends at the trypomastigote stage, but at the same alternative state as in the previous experiment. This indicates that these network perturbations could prevent the normal development of the parasite's life cycle.

Discussion

One fundamental open question in systems biology is how cells that share the same genome exhibit notably different gene-expression patterns or distinct phenotypes. This question is closely related to the process of establishing cell fates during development. A widely used picture to describe these phenomena is Waddington's epigenetic landscape, a phenomenological metaphor which corresponds to an energy landscape with many local minimums where the system moves regardless of whether environmental cues are present [36–38]. Despite the simplicity and elegance of Waddington's concept, it lacks quantitative mechanistic details. Given the significance of a quantitative understanding of cell phenotypic transitions, many efforts have been made to develop predictive mathematical frameworks [38–42]. Although some advances have been made for low dimensional systems, the application of these mathematical frameworks to higher dimensional models remains a theoretical challenge.

In this work, we have developed a reverse engineering approach to identify the gene network structures responsible for the observed dynamical properties of a high dimensional biological system. These dynamical properties include the steady states associated with the stable phenotypes, and the phenotypic transitions observed in *T. cruzi*'s life cycle. We have assumed that each of the five phenotypic transitions occurs in response to the external cue corresponding to the final state of the transition. For this reason, we have modeled the life cycle of *T. cruzi* as if it were an open dynamical system. Our methodology for embedding the observed expression patterns into the GRN dynamics adds several new ingredients such as the use of an ensemble of noise-perturbed training sets, and a pruning procedure to identify the significant network links. The information of the transitions between the stable phenotypes was used to develop an optimisation procedure. This reverse engineering procedure has been successfully used to identify one key network module that explains three of the five phenotypic transitions. To incorporate the information about phenotypic transitions we have assumed that all transitions occur at the same rate, and through the shortest path. Until now, there are not experimental facts supporting or not these assumptions. Thus, the present results must be interpreted taking into account the limitations of the available data. However, we believe that using our coarse-grain hypothesis about phenotypic transitions has more predictive power than only using the steady state transcriptome data. We are persuaded that having the transcriptome data between transitions could be valuable to improve our results.

In a previous work an interactome network of the early *T. cruzi* infection process was presented [18]. It was built using components of the ECM (THBS1, LAMC1, LGALS3, and ERK1/2) as initial seed nodes. *T. cruzi* gp83 ligand, a TS expressed only in invasive trypomastigotes, triggers gp83 receptors in the host cells via activation of ERK1/2. This results in the up-regulation of LAMC1 which, in turn, cross-talks with LGALS3 and THBS1. All these interactions enhance cellular infection using specific parasite surface molecules such as calreticulin (TcCRT), Tc45 mucin, and Tc85 [18]. In particular, infective trypomastigotes use Tc85, a member of the gp85/TS gene family, to interact with laminin [43], Tc45 mucin to interact with LAMC1 through LGALS3 [44], and TcCRT to interact with THBS1 [45]. Taken together, these are clear evidences that *T. cruzi* regulates and uses the ECM to invade host cells and cause

disease. In a certain way, our results are consistent with that, since some of the regulatory genes we have found code for virulence factors such as TSs or mucins. In addition, and regarding transitions, we have found 3 TS-coding genes (Q4CQC9, Q4DWU9 and Q4DLE3) and 1 mucin TcMUCII-coding gene (Q4D2K9) up-regulated in transitions leading to trypomastigote stage. This is another result that is in agreement with the previously mentioned works, in which the importance of certain parasite's molecules in parasite-host interactions and host-cell invasion was demonstrated.

Besides the development of a model with the ability to emulate the parasite's dynamics, the information presented in this work (S5 and S6 Tables) could be useful to assign previously unknown putative functions to some genes. In this sense, our results suggest that amastin genes could act as key regulators. This finding is consistent with a previous study in which it is shown that amastin may increase *T. cruzi*'s differentiation rates both in the insect and in the mammalian hosts [46]. On the other hand, our finding of TS-coding genes acting as regulators in the amastigote stage adds relevant information to the resulting parasite state. Furthermore, we have found that these same TS-coding genes are inhibited in the transitions leading to the amastigote stage, and activated in the transitions leading to the trypomastigote stage. Considering that TS plays a key role in *T. cruzi*'s infectivity and that this enzyme is not present in mammals, TS constitutes a potential target for the development of novel drugs to treat or prevent Chagas disease [47–49]. Finally, we have found four mucin genes belonging to the TcMUC family that act as regulators both in the mammalian and in the insect parasite's stages. It is known that this family of mucins is expressed only in the mammalian stages [50].

The present approach could be adapted to and useful for better understanding other single cell parasites with multiple developmental stages such as *T. brucei*, *P. falciparum* and *Leishmania*. Uncovering the core circuit that underlies the dynamics of these parasites' life cycles could open the door to new possibilities: the development of applications to reprogramme the parasites' life cycles, and the finding of new therapeutic targets against the parasites.

Supporting Information

S1 Fig. Clustering and dimension reduction. Davies-Bouldin index (DBI) as a function of the number of clusters, N_c, used in the clustering procedure. The arrow in $N_c = 339$ indicates the optimal number of clusters used in subsequent procedures.
(TIF)

S2 Fig. Main regulatory clusters for each steady state. The plots represent the networks derived from amastigote (a), epimastigote (b), metacyclic tryp. (c), and trypomastigote (d) stages.
(TIF)

S3 Fig. Network dynamics' *in silico* perturbations. Yellow trajectory shows the dynamics of the system, in the space spanned by the three principal components, when steady state corresponding to the trypomastigote stage is perturbed by mean of over-expressing and knocking-down genes belonging to cluster 326 and 337, respectively. Blue trajectory shows the dynamics of the system when clusters 259, 260, and 332, predicted to be down-regulated by the external cue leading to the trypomastigote stage, are overexpressed. For comparison, trajectories corresponding to the unperturbed transition from amastigote to trypomastigote stages are also plotted (colored circles).
(TIF)

S1 Movie. Transition from the amastigote stage to the trypomastigote stage. The animated matrix plots show the evolution of the system from the amastigote stage, under the action of $\mu = 4$, to the trypomastigote stage. The movie is composed of 12 frames; one for each step in the

simulation. The simulation shows a clear similarity between the states associated with the last two frames and the corresponding target stage indicated by μ.
(AVI)

S2 Movie. Transition from the trypomastigote stage to the epimastigote stage. The animated matrix plots show the evolution of the system from the trypomastigote stage, under the action of $\mu = 2$, to the epimastigote stage. The movie is composed of 12 frames; one for each step in the simulation. The simulation shows a clear similarity between the states associated with the last two frames and the corresponding target stage indicated by μ.
(AVI)

S3 Movie. Transition from the epimastigote stage to the metacyclic tryp. stage. The animated matrix plots show the evolution of the system from the epimastigote stage, under the action of $\mu = 3$, to the metacyclic tryp. stage. The movie is composed of 12 frames; one for each step in the simulation. The simulation shows a clear similarity between the states associated with the last two frames and the corresponding target stage indicated by μ.
(AVI)

S4 Movie. Transition from the metacyclic tryp. stage to the amastigote stage. The animated matrix plots show the evolution of the system from the metacyclic tryp. stage, under the action of $\mu = 1$, to the amastigote stage. The movie is composed of 12 frames; one for each step in the simulation. The simulation shows a clear similarity between the states associated with the last two frames and the corresponding target stage indicated by μ.
(AVI)

S5 Movie. Transition from the trypomastigote stage to the amastigote stage. The animated matrix plots show the evolution of the system from the trypomastigote stage, under the action of $\mu = 1$, to the amastigote stage. The movie is composed of 12 frames; one for each step in the simulation. The simulation shows a clear similarity between the states associated with the last two frames and the corresponding target stage indicated by μ.
(AVI)

S1 Table. Gene-expression profile of *T. cruzi*'s life cycle. Log-norm expression levels corresponding to 8,904 *T. cruzi* genes, obtained from microarray experiments as indicated in Methods. The gene IDs listed in the first column correspond to our own gene numbering. The second column lists the microarray oligo IDs. The third and the fourth columns list the gene and protein names, respectively. The last four columns correspond to the gene-expression levels in each stage of the parasite's life cycle: amastigote, epimastigote, metacyclic tryp. and trypomastigote, respectively.
(XLSX)

S2 Table. Cluster composition. Clusters (cluster ID) are listed in the first column, with their corresponding genes (gene ID) listed in the second column. The third and the fourth columns list the gene and protein names, respectively.
(XLSX)

S3 Table. Intra-cluster averages of the expression levels. Gene-expression levels of each cluster (rows) in each different stage (columns) used in all further modeling computations.
(XLSX)

S4 Table. Regulatory links. List of the 11,470 significant weights (at a significance level of 0.01) needed for the maintenance of the parasite's steady states. The first column indicates the regulatory cluster IDs. The second column indicates the regulated cluster IDs. The third

column lists the mean values of each cluster, averaged over the ensemble of 300 training sets. The fourth column lists the associated standard deviations. The last column lists the *p*-values (the probabilities of the location test).
(XLSX)

S5 Table. Main regulatory genes. List of the genes with important regulatory activity for the maintenance of the parasite's steady states. The steady states are indicated by a capital letter: amastigote (A), epimastigote (E), metacyclic tryp. (M) and trypomastigote (T).
(XLSX)

S6 Table. Main regulated genes. List of the genes regulated by the external cues responsible for the transitions between the steady states.
(XLSX)

S7 Table. Analysis of uncharacterised proteins. Result of the Interproscan analysis of the uncharacterised proteins without GO annotations, listed in S5 and S6 Tables. Uncharacterised proteins with no Interproscan information are listed at the end.
(XLSX)

S8 Table. Regulatory links of *T. cruzi*'s life cycle subnetwork. Estimated values of parameters $w_{i,j}$, θ_i, and k_i^{μ}, extracted from matrix $\mathbf{W_t}$.
(XLSX)

Acknowledgments

The authors wish to thank Geoffrey Siwo for his critical review of the manuscript. A.C. is a postdoctoral fellow of the CONICET (Argentina). L.D. is a research member of the CONICET (Argentina).

Author Contributions

Conceived and designed the experiments: LD. Performed the experiments: AC LD. Analyzed the data: AC LD. Wrote the paper: AC LD.

References

1. Huang S, Guo YP, Enver T, May G. Bifurcation dynamics in lineage-commitment in bipotent progenitor cells. Developmental Biology. 2007; 305(2):695–713. doi: 10.1016/j.ydbio.2007.02.036 PMID: 17412320
2. Macarthur BD, Ma'ayan A, Lemischka IR. Systems biology of stem cell fate and cellular reprogramming. Nature Reviews Molecular Cell Biology. 2009; 10(10):672–681. doi: 10.1038/nrm2766 PMID: 19738627
3. Henry A, Monéger F, Samal A, Martin OC. Network function shapes network structure: the case of the Arabidopsis flower organ specification genetic network. Molecular BioSystems. 2013; 9(7):1726–1735. doi: 10.1039/c3mb25562j PMID: 23579205
4. Zhou JX, Brusch L, Huang S. Predicting pancreas cell fate decisions and reprogramming with a hierarchical multi-attractor model. PLoS ONE. 2011; 6(3):e14752. doi: 10.1371/journal.pone.0014752 PMID: 21423725
5. Lang AH, Li H, Collins JJ, Mehta P. Epigenetic landscapes explain partially reprogrammed cells and identify key reprogramming genes. PLoS Computational Biology. 2014 08; 10(8):e1003734. doi: 10.1371/journal.pcbi.1003734 PMID: 25122086
6. Basso K, Margolin A, Stolovitzky G, Klein U, Dalla-Favera R, Califano A. Reverse engineering of regulatory networks in human B cells. Nature Genetics. 2005 Apr; 37(4):382–390. doi: 10.1038/ng1532 PMID: 15778709

7. Novershtern N, Subramanian A, Lawton LN, Mak RH, Haining WN, McConkey ME, et al. Densely interconnected transcriptional circuits control cell states in human hematopoiesis. Cell. 2011; 144:296–309. doi: 10.1016/j.cell.2011.01.004 PMID: 21241896
8. Moignard V, Macaulay IC, Swiers G, Buettner F, Schütte J, Calero-Nieto FJ, et al. Characterization of transcriptional networks in blood stem and progenitor cells using high-throughput single-cell gene expression analysis. Nature Cell Biology. 2013; 15(4):363–372. doi: 10.1038/ncb2709 PMID: 23524953
9. Marbach D, Prill RJ, Schaffter T, Mattiussi C, Floreano D, Stolovitzky G. Revealing strengths and weaknesses of methods for gene network inference. Proceedings of the National Academy of Sciences of the USA. 2010; 107(14):6286–6291. doi: 10.1073/pnas.0913357107 PMID: 20308593
10. Molinelli EJ, Korkut A, Wang W, Miller ML, Gauthier NP, Jing X, et al. Perturbation biology: inferring signaling networks in cellular systems. PLoS Computational Biology. 2013; 9(12):e1003290. doi: 10.1371/journal.pcbi.1003290 PMID: 24367245
11. Margolin A, Califano A. Theory and limitations of genetic network inference from microarray data. Annals of the New York Academy of Sciences. 2007; 1115:51–72. doi: 10.1196/annals.1407.019 PMID: 17925348
12. World Health Organization. Fact sheet N°340; 2015.
13. Rassi A Jr, Rassi A, Marin-Neto JA. Chagas disease. Lancet. 2010; 375(9723):1388–1402. doi: 10.1016/S0140-6736(10)60061-X
14. Rassi A, Marcondes de Rezende J. American trypanosomiasis (Chagas disease). Infectious disease clinics of North America. 2012; 26(2):275–291. doi: 10.1016/j.idc.2012.03.002 PMID: 22632639
15. Gupta S, Garg NJ. TcVac3 induced control of Trypanosoma cruzi infection and chronic myocarditis in mice. PLoS ONE. 2013; 8:e59434. doi: 10.1371/journal.pone.0059434 PMID: 23555672
16. Minning T, Weatherly DB, Atwood J, Orlando R, Tarleton RL. The steady-state transcriptome of the four major life-cycle stages of Trypanosoma cruzi. BMC Genomics. 2009; 10:370. doi: 10.1186/1471-2164-10-370 PMID: 19664227
17. Nde PN, Johnson CA, Pratap S, Cardenas TC, Kleshchenko YY, Furtak VA, et al. Gene network analysis during early infection of human coronary artery smooth muscle cells by Trypanosoma cruzi and Its gp83 ligand. Chemistry & biodiversity. 2010; 7(5):1051–1064. doi: 10.1002/cbdv.200900320
18. Nde PN, Lima MF, Johnson CA, Pratap S, Villalta F. Regulation and use of the extracellular matrix by Trypanosoma cruzi during early infection. Frontiers in Immunology. 2012; 3:1–10. doi: 10.3389/fimmu.2012.00337
19. Mar JC, Wells CA, Quackenbush J. Defining an informativeness metric for clustering gene expression data. Bioinformatics. 2011; 27:1094–1100. doi: 10.1093/bioinformatics/btr074 PMID: 21330289
20. Diambra L. Clustering gene expression by dynamics: A maximum entropy approach. Physica A. 2008; 387:2187–2196. doi: 10.1016/j.physa.2007.12.006
21. Davies DL, Bouldin DW. A cluster separation measure. IEEE transactions on pattern analysis and machine intelligence. 1979; 1:224–227. doi: 10.1109/TPAMI.1979.4766909 PMID: 21868852
22. D'haeseleer P, Wen X, Fuhrman S, Somogyi R. Linear modeling of mRNA expression levels during CNS development and injury. Pacific Symposium on Biocomputing. 1999; 4:41–52.
23. Diambra L. Coarse-grain reconstruction of genetic networks from expression levels. Physica A. 2011 Jun; 390(11):2198–2207. doi: 10.1016/j.physa.2011.02.021
24. Michailidis G, D'Alché-Buc F. Autoregressive models for gene regulatory network inference: sparsity, stability and causality issues. Mathematical Biosciences. 2013; 246(2):326–334. doi: 10.1016/j.mbs.2013.10.003 PMID: 24176667
25. Yeung MS, Tegner J, Collins J. Reverse engineering gene networks using singular value decomposition and robust regression. Proceedings of the National Academy of Sciences of the USA. 2002; 99(9):6163–6168. doi: 10.1073/pnas.092576199 PMID: 11983907
26. Alter O, Brown PO, Botstein D. Singular value decomposition for genome-wide expression data processing and modeling. Proceedings of the National Academy of Sciences of the USA. 2000; 97(18):10101–10106. doi: 10.1073/pnas.97.18.10101 PMID: 10963673
27. Bishop CM. Training with noise is equivalent to Tikhonov regularization. Neural Computation. 1995 Jan; 7(1):108–116. doi: 10.1162/neco.1995.7.1.108
28. Zur RM, Jiang Y, Pesce LL, Drukker K. Noise injection for training artificial neural networks: A comparison with weight decay and early stopping. Medical Physics. 2009; 36(10):4810–4818. doi: 10.1118/1.3213517 PMID: 19928111
29. Good PI, Hardin JW. Common Errors in Statistics (and How to Avoid Them), Third Edition; 2008.
30. Friedman N, Linial M, Nachman I, Pe'er D. Using Bayesian networks to analyze expression data. J Comput Biol. 2000; 7:601–620. doi: 10.1089/106652700750050961 PMID: 11108481

31. Akutsu T, Miyano S, Kuhara S. Identification of genetic networks from a small number of gene expression patterns under the Boolean network model. Pacific Symposium on Biocomputing. 1999; 4:17–28. PMID: 10380182

32. D'haeseleer P, Liang S, Somogyi R. Genetic network inference: from co-expression clustering to reverse engineering. Bioinformatics. 2000; 16:707–726. doi: 10.1093/bioinformatics/16.8.707 PMID: 11099257

33. Thieffry D, Huerta AM, Pérez-Rueda E, Collado-vides J. From specific gene regulation to genomic networks: a global analysis of transcriptional regulation in Escherichia coli. BioEssays. 1998; 20(5):433–440. doi: 10.1002/(SICI)1521-1878(199805)20:5%3C433::AID-BIES10%3E3.0.CO;2-2 PMID: 9670816

34. Farkas I, Jeong H, Vicsek T, Barabási AL, Oltvaia Z, Oltvai ZN. The topology of the transcription regulatory network in the yeast, Saccharomyces cerevisiae. Physica A. 2003; 318:601–612. doi: 10.1016/S0378-4371(02)01731-4

35. Jones P, Binns D, Chang HY, Fraser M, Li W, McAnulla C, et al. InterProScan 5: Genome-scale protein function classification. Bioinformatics. 2014; 30(9):1236–1240. doi: 10.1093/bioinformatics/btu031 PMID: 24451626

36. Frauenfelder H, Sligar SG, Wolynes PG. The energy landscapes and motions of proteins. Science. 1991; 254:1598–1603. PMID: 1749933

37. Zhou JX, Huang S. Understanding gene circuits at cell-fate branch points for rational cell reprogramming. Trends in Genetics. 2011 Feb; 27(2):55–62. doi: 10.1016/j.tig.2010.11.002 PMID: 21146896

38. Ferrell JE. Bistability, bifurcations, and Waddington's epigenetic landscape. Current Biology. 2012 Jun; 22(11):R458–R466. doi: 10.1016/j.cub.2012.03.045 PMID: 22677291

39. Wang J, Li C, Wang E. Potential and flux landscapes quantify the stability and robustness of budding yeast cell cycle network. Proceedings of the National Academy of Sciences of the USA. 2010; 107:8195–8200. doi: 10.1073/pnas.0910331107 PMID: 20393126

40. Wang J, Zhang K, Xu L, Wang E. Quantifying the Waddington landscape and biological paths for development and differentiation. Proceedings of the National Academy of Sciences of the USA. 2011; 108:8257–8262. doi: 10.1073/pnas.1017017108 PMID: 21536909

41. Li C, Wang J. Quantifying Waddington landscapes and paths of non-adiabatic cell fate decisions for differentiation, reprogramming and transdifferentiation. Journal of the Royal Society, Interface. 2013; 10:20130787. doi: 10.1098/rsif.2013.0787 PMID: 24132204

42. Li C, Wang J. Quantifying cell fate decisions for differentiation and reprogramming of a human stem cell network: Landscape and biological paths. PLoS Computational Biology. 2013; 9:e1003165. doi: 10.1371/journal.pcbi.1003165 PMID: 23935477

43. Giordano R, Fouts DL, Tewari D, Colli W, Manning JE, Alves MJM. Cloning of a surface membrane glycoprotein specific for the infective form of Trypanosoma cruzi having adhesive properties to laminin. Journal of Biological Chemistry. 1999; 274(6):3461–3468. doi: 10.1074/jbc.274.6.3461 PMID: 9920891

44. Moody TN, Ochieng J, Villalta F. Novel mechanism that Trypanosoma cruzi uses to adhere to the extracellular matrix mediated by human galectin-3. FEBS Letters. 2000; 470(3):305–308. doi: 10.1016/S0014-5793(00)01347-8 PMID: 10745086

45. Johnson CA, Kleshchenko YY, Ikejiani AO, Udoko AN, Cardenas TC, Pratap S, et al. Thrombospondin-1 interacts with Trypanosoma cruzi surface calreticulin to enhance cellular infection. PLoS ONE. 2012; 7(7):e40614. doi: 10.1371/journal.pone.0040614 PMID: 22808206

46. Cruz MC, Souza-Melo N, Vieira da Silva C, Duarte daRocha W, Bahia D, Araújo PR, et al. Trypanosoma cruzi: Role of δ-amastin on extracellular amastigote cell invasion and differentiation. PLoS ONE. 2012; 7(12):e51804. doi: 10.1371/journal.pone.0051804 PMID: 23272170

47. Neres J, Brewer ML, Ratier L, Botti H, Buschiazzo A, Edwards PN, et al. Discovery of novel inhibitors of Trypanosoma cruzi trans-sialidase from in silico screening. Bioorganic and Medicinal Chemistry Letters. 2009; 19(3):589–596. doi: 10.1016/j.bmcl.2008.12.065 PMID: 19144516

48. Carvalho ST, Sola-Penna M, Oliveira IA, Pita S, Gonçalves AS, Neves BC, et al. A new class of mechanism-based inhibitors for trypanosoma cruzi trans-sialidase and their influence on parasite virulence. Glycobiology. 2010; 20(8):1034–1045. doi: 10.1093/glycob/cwq065 PMID: 20466651

49. Buschiazzo A, Muiá R, Larrieux N, Pitcovsky T, Mucci J, Campetella O. Trypanosoma cruzi trans-sialidase in complex with a neutralizing antibody: Structure/function studies towards the rational design of inhibitors. PLoS Pathogens. 2012; 8(1):e1002474. doi: 10.1371/journal.ppat.1002474 PMID: 22241998

50. Buscaglia CA, Campo VA, Frasch AC, Di Noia JM. Trypanosoma cruzi surface mucins: Host-dependent coat diversity. Nature Reviews Microbiology. 2006; 4(3):229–236. doi: 10.1038/nrmicro1351 PMID: 16489349