

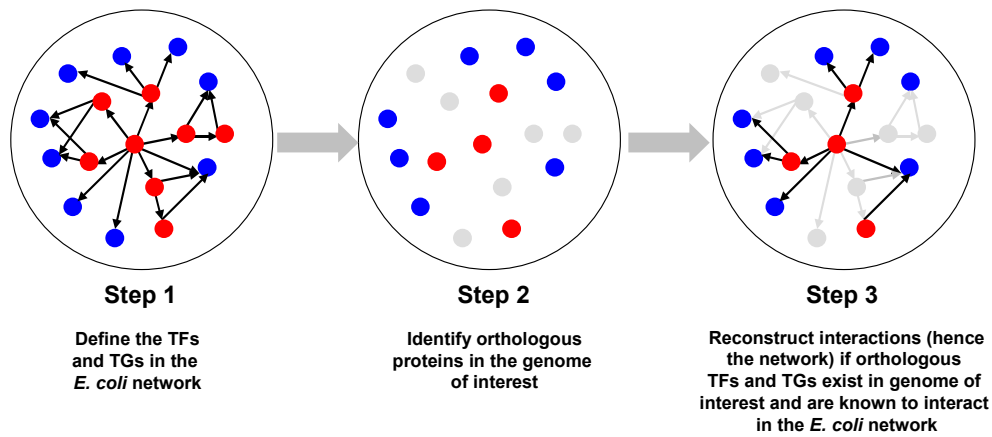
# SUPPLEMENTARY METHODS

M1: ALGORITHM TO RECONSTRUCT TRANSCRIPTIONAL NETWORKS .....	M-2
<i>Figure 1: Procedure to reconstruct transcriptional regulatory networks .....</i>	<i>M-2</i>
M2: PROCEDURE TO IDENTIFY ORTHOLOGOUS PROTEINS .....	M-3
<i>Figure 2: Flowchart describing the procedure to detect orthologs .....</i>	<i>M-3</i>
<i>Figure 3: Schematic of the procedure to detect orthologous proteins .....</i>	<i>M-4</i>
M3: PROCEDURE TO EVALUATE SIGNIFICANCE OF THE BIAS IN GENE CONSERVATION.....	M-5
<i>Figure 4: Procedure to create random networks to evaluate TF and TG conservation .....</i>	<i>M-5</i>
M4: ALGORITHM TO RECONSTRUCT ANCESTRAL NETWORKS .....	M-6
<i>Figure 5: Species tree.....</i>	<i>M-6</i>
<i>Figure 6: The Dollo approach taken to calculate ancestral networks.....</i>	<i>M-6</i>
M5: PROCEDURE TO GROUP GENOMES BY INTERACTIONS AND GENES CONSERVED.....	M-7
<i>Figure 7: Procedure to cluster genomes based on interactions conserved .....</i>	<i>M-7</i>
M6: ALGORITHM TO SIMULATE NETWORK EVOLUTION .....	M-8
M7: PROCEDURE TO ANALYSE SCALE FREE BEHAVIOUR OF CONSERVED NETWORKS .....	M-9
<i>Figure 8: Procedure to create random networks as seen in the genome of interest .....</i>	<i>M-9</i>
M8: ALGORITHM TO ANALYSE CONSERVATION OF EQUIVALENT ‘NETWORK MOTIFS’ .....	M-10
<i>Figure 9: Procedure to cluster genomes and motifs according to the extent conserved.....</i>	<i>M-10</i>
M9: PROCEDURE TO EVALUATE SIGNIFICANCE OF MOTIF INTERACTION CONSERVATION .....	M-11
M10: PROCEDURE TO EVALUATE SIGNIFICANCE OF MOTIF INTERACTION CONSERVATION .....	M-13

## M1: Algorithm to reconstruct transcriptional networks

The transcriptional regulatory network for *E. coli* was used as the basis to reconstruct networks for other genomes. Information about regulatory interactions was obtained from RegulonDB (Salgado et al., 2004) and the dataset used in Shen-Orr *et al.* (Shen-Orr et al., 2002) This provided us with 1295 transcriptional interactions involving 755 proteins (112 transcription factors). Orthologous proteins were identified in the genome of interest. If orthologs were identified for an interacting transcription factor and target gene in *E. coli*, then an interaction was considered to be present in the genome of interest (**Figure 1**).

**Figure 1: Procedure to reconstruct transcriptional regulatory networks**



**Figure 1:** This figure illustrates the steps involved in reconstructing regulatory networks for the genomes. Blue circles represent target genes and red circle represents transcription factors. Black arrow represents a transcriptional interaction. Genes and interactions that are absent are shown in grey. Reconstructed transcriptional regulatory networks for the 175 genomes using the method discussed above are available from the supplementary material website at <http://www.mrc-lmb.cam.ac.uk/genomes/madanm/evdy/>.

## M2: Procedure to identify orthologous proteins

Detecting orthology is non-trivial. After testing various orthology detection procedures (bi-directional best hit, best hits with defined e-value cut-offs, etc), we arrived at a hybrid procedure, which was used to identify orthologous proteins in a genome. Bi-directional best hit is a very conservative approach to detect orthologs. It performs best for closely related organisms and may fail to pick orthologs from distantly related organisms. Best hit method using specific cut-offs is too liberal, and may result in false positive hits when the genomes compared are distantly related. So our hybrid method uses a combination of two methods as described in a flow chart below (Figure 2):

Figure 2: Flowchart describing the procedure to detect orthologs

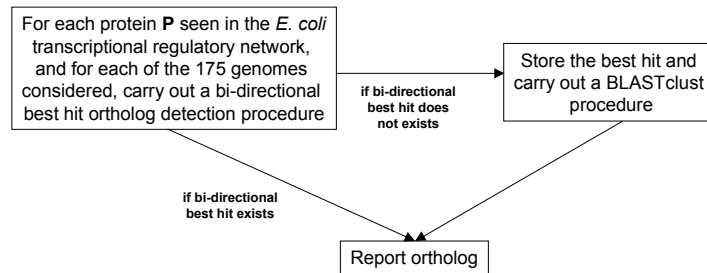


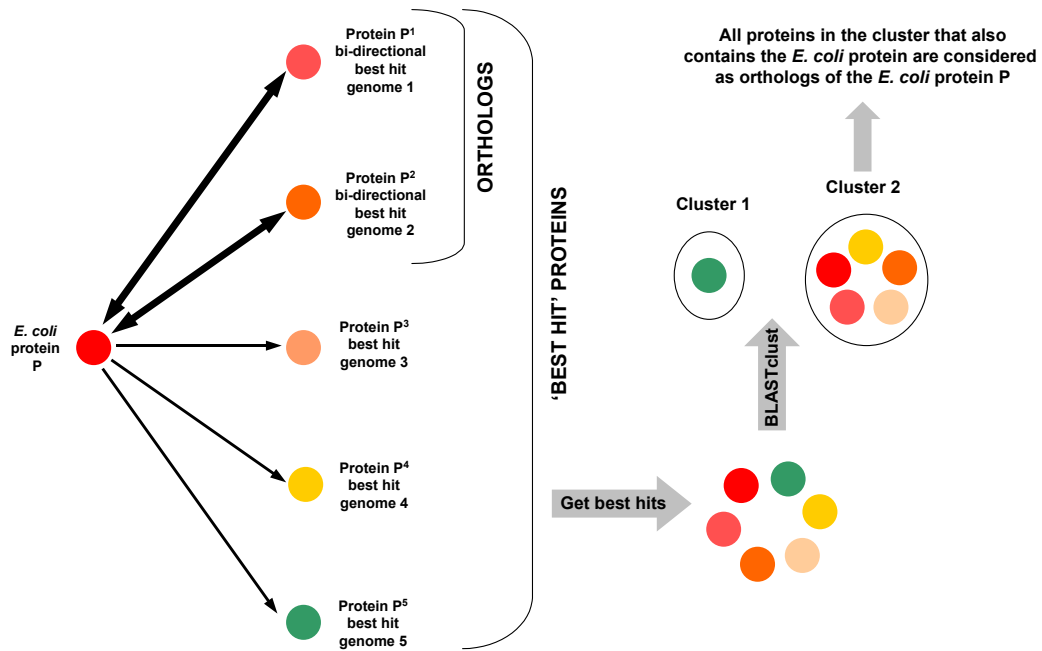
Figure 2: First a bi-directional best hit is performed, if an ortholog is not detected then BLASTclust procedure is adopted.

**Bi-directional best-hit procedure:** For each protein  $P$  in the *E. coli* network, a BLAST search was performed against the genome of interest ( $x$ ). The best hit, sequence  $P^x$  from genome  $x$ , was then used as a query and a BLAST search was carried out against the *E. coli* genome. If the best hit using  $P^x$  as the query happens to be  $P$  in *E. coli*, then  $P$  &  $P^x$  were considered as orthologous proteins. If however  $P^x$  does not pick up  $P$  from *E. coli* as its best hit, then a BLASTclust (Lespinet et al., 2002) procedure was adopted.

**BLASTclust algorithm:** For each of the proteins  $P$  in *E. coli* for which the above-mentioned procedure did not pick up orthologous proteins, the best-hit sequences  $P^x$  (using  $P$  as the query against genome  $x$ ) for each genome were obtained. Thus, for every protein  $P$ , this procedure gives us a set of proteins, which are the best hits from genomes where bi-directional best-hit procedure fails. The set of sequences thus obtained along with the *E. coli* protein is taken through a BLASTclust procedure using length conservation ( $L$ ) of 60% and a score density ( $S$ ) of 60% (Lespinet et al., 2002). Score density ( $S$ ) is defined as the ratio of number of identical residues in the alignment to the length of the alignment. Documentation is available at: <http://bio.ifom-firc.it/docs/blast/readme.bcl.txt> and in (Lespinet et al., 2002). This procedure first carries out an all against all sequence comparison and produces clusters of sequences using the single linkage-clustering algorithm. This step of all against all comparison will ensure that only orthologous proteins in distantly related organisms will still be picked up reliably through the sequences from intermediately distant genomes. All the sequences belonging to the cluster that also contains the *E. coli* protein  $P$  are considered as orthologs (see schematic of the ortholog detection procedure). Analysis of the clusters using various combination of procedures reveal that the parameters  $S=0.6$  and  $L=0.6$  performs best with an optimum coverage and lowest false positive rate (Lespinet et al., 2002). See Figure 3 for a schematic.

This combination method, of the conservative bi-directional best hit with the blastclust procedure that can detect orthologs from distantly related organisms through intermediate sequences was used to detect orthologs reliably.

**Figure 3: Schematic of the procedure to detect orthologous proteins**



**Figure 3:** This figure illustrates the steps involved in ortholog detection. Information about orthologs for every genome is stored as an 'orthomap' file and is available from the as supplementary material website.

### M3: Procedure to evaluate significance of the bias in gene conservation

Analysis of the conservation of transcription factors and target genes for the various genomes indicate that target genes are more conserved than transcription factors. To evaluate the significance of this observation, the following procedure was carried out. For each of the 175 genomes, orthologous proteins in the *E. coli* network were first identified. Among the identified orthologs, the fraction of TFs and TGs conserved were calculated. A graph of %TFs conserved against % TGs conserved was plotted. The slope and intercept of the best fit to the observed trend was obtained. Then for each of the genome, random networks of similar sizes were created (Figure 4) 10,000 times. For each run, the slope and intercept for the plot of % TF conserved against % TG conserved was obtained. The P-value was estimated as the number of runs where the slope was less than or equal to what was observed.

Figure 4: Procedure to create random networks to evaluate TF and TG conservation

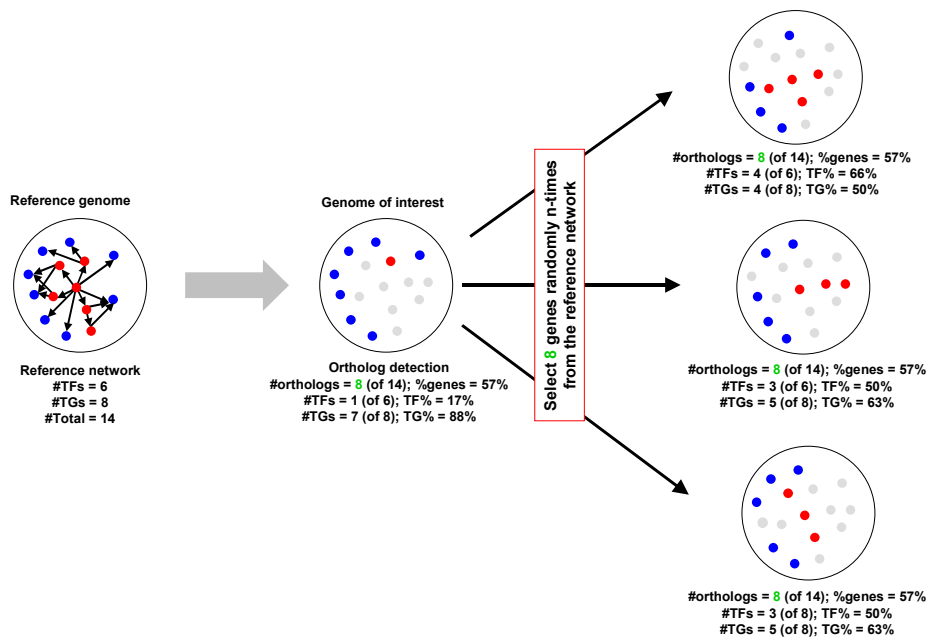


Figure 4: This figure illustrates the procedure to generate random networks of similar size to evaluate significance of TF and TG conservation in the different genomes.

#### M4: Algorithm to reconstruct ancestral networks

Standard species tree (Figure 5) was adapted from Margulis and Schwartz (4), as shown in the figure below. Only organisms whose genome size was greater than half the size of the *E. coli* genome were considered to avoid any bias arising from parasitic genomes that have lost genes due to a specialised life style. This criterion resulted in a total of 97 genomes.

Figure 5: Species tree

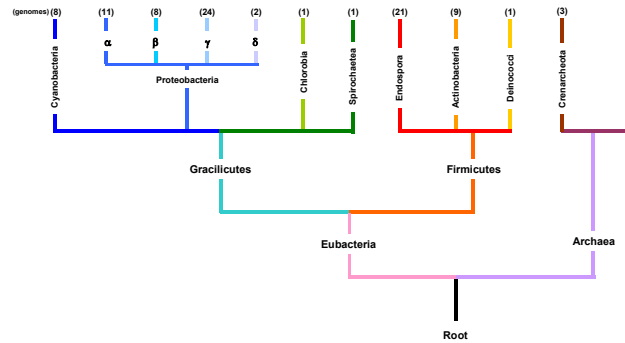


Figure 5: Species tree adopted for calculating ancestral networks.

Figure 6: The Dollo approach taken to calculate ancestral networks

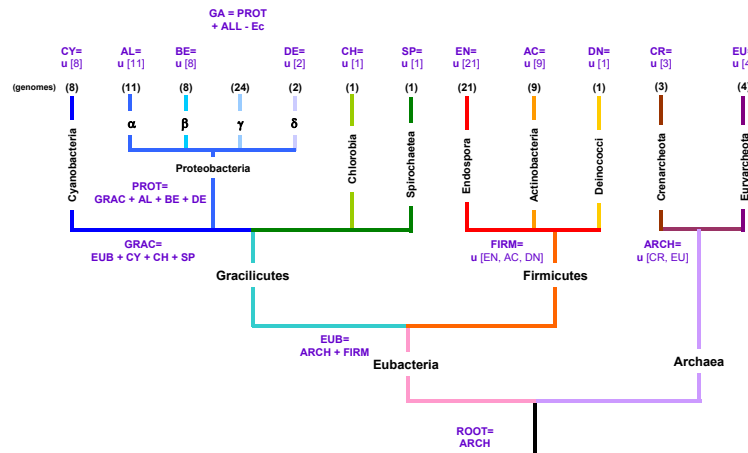


Figure 6: The figure illustrates the principle to calculate the ancestral networks using the Dollo parsimony approach.

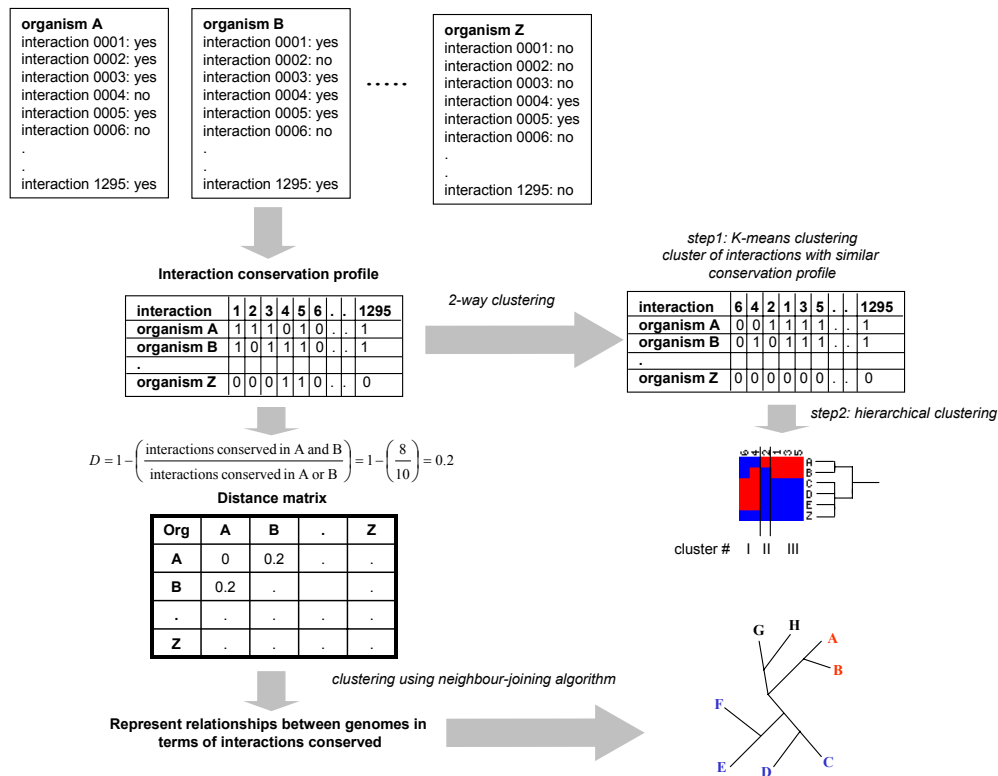
A gene is considered to be present in the node if at least one of its child nodes contains it. The function U represents the union operation. For example, the node  $CY = u [8]$  means that the ancestor for all cyanobacterial genomes (8 genomes in this case) will contain all the genes seen in each of the 8 cyanobacterial genomes. The Dollo principle was used to reconstruct the ancestral states for all the ancestral nodes. This principle states that a gene (ortholog) can be gained only once. The emergence of a gene was thus assigned to the last common ancestor of all lineages that contained the given gene (Figure 6). Once the gene composition has been determined for each node in the tree, the transcriptional regulatory network was reconstructed as described in the procedure mentioned before.

## M5: Procedure to group genomes by interactions and genes conserved

A novel method to analyse conservation of interaction was developed (Figure 7). For every genome, an interaction conservation profile was calculated based on the reconstructed networks. A distance matrix based on interactions conserved was calculated. Relationships between genomes, and within interactions/genes were represented as a tree and a matrix using the procedure shown on the next page.

This procedure was carried out for genes instead of transcriptional interactions. Using this procedure, clusters of transcription factors that have similar conservation profiles were obtained.

**Figure 7: Procedure to cluster genomes based on interactions conserved**



**Figure 7: Representing the interactions as a vector (interaction profile) makes the network amenable to standard clustering procedure.**

## **M6: Algorithm to simulate network evolution**

Since there was a bias in the conservation of target genes and transcription factors, we were interested in asking if there was a selection for regulatory hubs to be conserved or not. The table in the main text shows that there is no trend for regulatory hubs to be conserved. To see how it affects the network structure in terms of the interactions being conserved, the following simulation procedure was carried out.

(a) Selection for hubs:

- 01:** order transcription factors (ascending) according to the # of target genes they regulate
- 02:** for each of the 175 genomes
- 03:** identify # of TF and # of TG
- 04:** create network with same # of TF and # of TG for genome, but choosing the TF with highest connectivity first
- 05:** reconstruct transcriptional interactions
- 06:** plot % interactions lost v/s % genes conserved

(b) Neutral conservation of genes:

- 01:** for each of the 175 genomes
- 02:** identify # of TF and # of TG
- 03:** create network with same # of TF and # of TG for genome, choosing TF & TG without bias
- 04:** reconstruct transcriptional interactions
- 05:** plot % interactions lost v/s % genes conserved

(c) Selection against hubs:

- 01:** order transcription factors (descending) according to the # of target genes they regulate
- 02:** for each of the 175 genomes
- 03:** identify # of TF and # of TG
- 04:** create network with same # of TF and # of TG for genome, but choosing the TF with the lowest connectivity first
- 05:** reconstruct transcriptional interactions
- 06:** plot % interactions lost v/s % genes conserved

(d) Random conservation of non-interacting pairs of genes:

- 01:** for each of the 175 genomes
- 02:** identify # of interactions in reconstructed network
- 03:** create network with same number of interactions but with pairs that are known not to interact – it can be TG-TG pair or TF-TG, TF-TG non-interacting pair
- 04:** plot % interactions lost v/s % genes conserved

## M7: Procedure to analyse scale free behaviour of conserved networks

The distribution of outgoing connectivity provides an indication about the large-scale structure of networks. It is well established that the outgoing connectivity for the *E. coli* network follows a scale-free behaviour, i.e. the distribution is best approximated by a power-law function  $T = aK^{-b}$  where  $T$  is the number of transcription factors with  $K$  connections. To evaluate the distribution for the reconstructed networks, the following procedure was carried out: For each of the 175 genomes, the distribution is approximated by a linear function ( $T = a + bK$ ), exponential function ( $T = ae^{-K\alpha}$ ;  $\log T = \log a - K\alpha \log e$ ) and a power-law function ( $T = aK^{-\gamma}$ ;  $\log T = \log a - \gamma \log K$ ). The function that best approximates the observed distribution is identified as the one that has the lowest error. To identify the trend in random networks the following procedure was carried out: for each of 10,000 times create 175 random networks of similar sizes to what is observed in the genomes (Figure 8). The procedure explained above is carried out to get the function that best approximates the distribution for the random networks. For each of the 17g genomes, the mean and standard deviation for the power-law exponent over the 10,000 runs is calculated. P-value was calculated as the fraction of the runs where the value for the exponent was greater than or equal to the observed value. Z-score was calculated as  $Z = (\gamma^{obs} - \gamma^{mean})/\sigma$ .

Figure 8: Procedure to create random networks as seen in the genome of interest

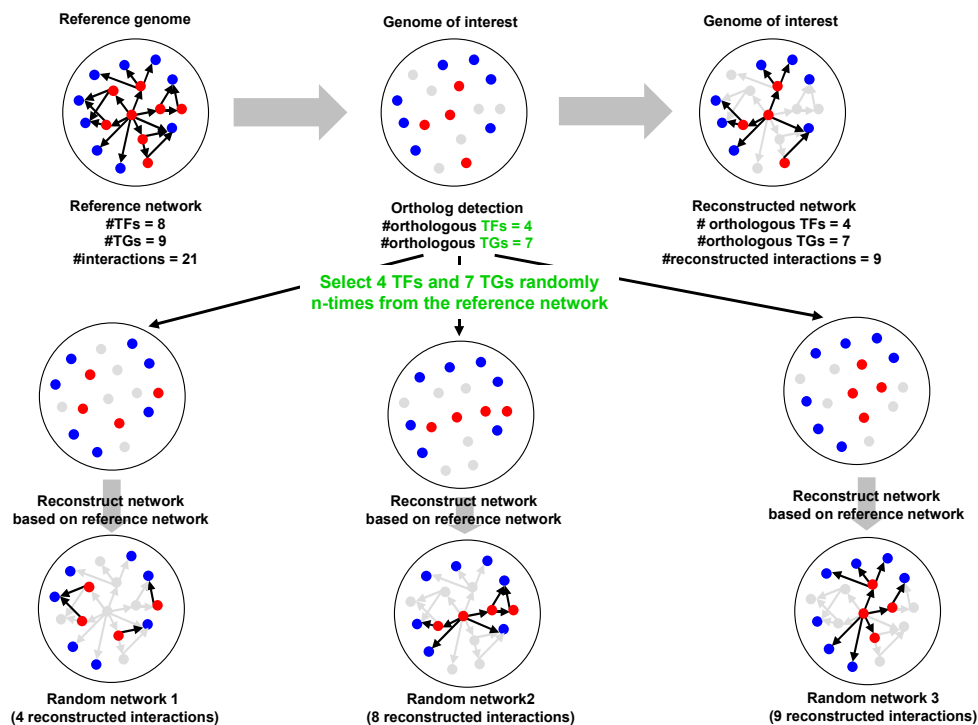
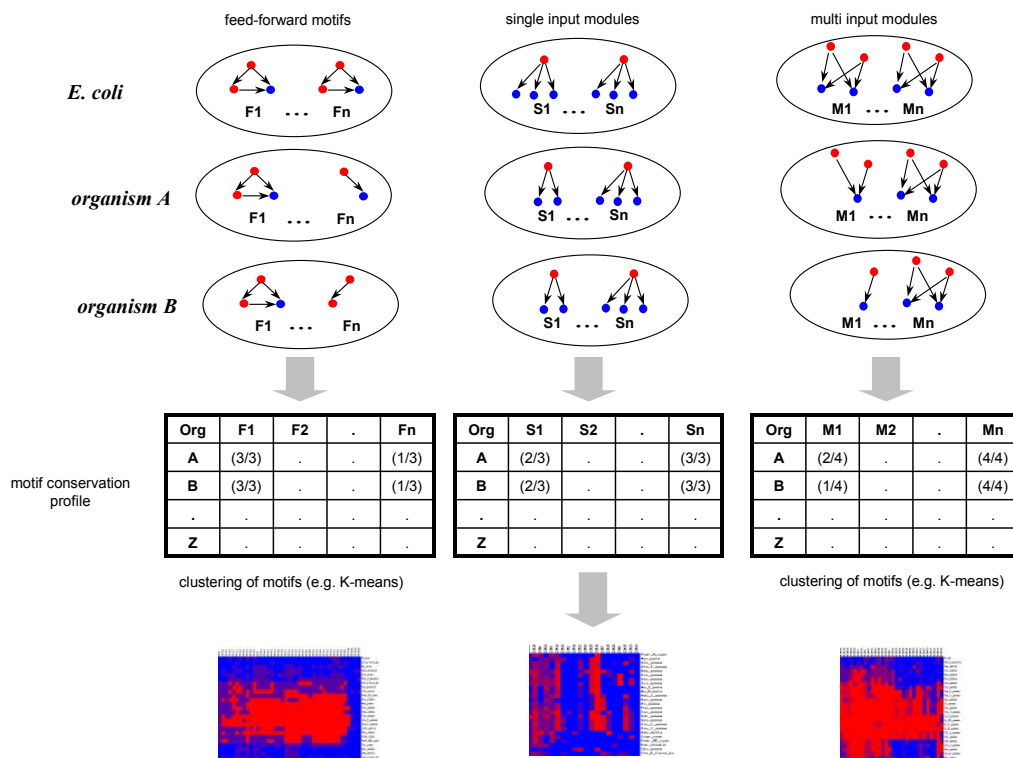


Figure 8: Random networks are generated by randomly choosing the same number of genes as seen in the genome of interest and interactions are reconstructed as discussed before.

### M8: Algorithm to analyse conservation of equivalent ‘network motifs’

A novel algorithm to analyse conservation of network motifs developed by us is shown in **Figure 9** below: Feed forward motif (FFM), single input module (SIM) and multi-input module (MIM) motifs were identified using the methods described by Shen-Orr *et al* (Shen-Orr *et al.*, 2002). A motif was considered to be absolutely conserved in a genome, if all the genes constituting the motif in *E. coli* were conserved in the other genome. If some genes were missing, the fraction of interactions in the motifs conserved was noted. Thus for each genome, an ordered n-dimensional vector (motif conservation profile) was created, where n is the number of motifs considered. The values represent the fraction of the interactions forming the motifs that are conserved. This matrix was then subjected to the procedure explained in section 4.2.5 to get motifs that have similar conservation profile.

**Figure 9: Procedure to cluster genomes and motifs according to the extent conserved**



**Figure 9:** The two way clustering provides information about which motifs have been completely conserved in set of genomes. Specific examples are discussed in the main text.

## M9: Procedure to evaluate significance of motif interaction conservation

To evaluate whether interactions in a motif are more conserved than any interaction in the network, we introduce a term called conservation index (C.I.), which is defined as follows:

$$C.I._{genome\ X} = \log_2 \left( \frac{R_{motif}}{R_{all}} \right)$$

$$R_{motif} = \frac{I_{genome\ X}^{motif}}{I_{E.coli}^{motif}} \quad \text{and} \quad R_{all} = \frac{I_{genome\ X}^{all}}{I_{E.coli}^{all}}$$

In this definition,  $I_{genome\ X}^{motif}$  is the number of interactions that forms a motif in *E. coli*, which has been conserved in genome X.  $I_{E.coli}^{motif}$  is the number of interactions in a motif in *E. coli*.  $I_{genome\ X}^{all}$  is the total number of interactions that have been conserved in genome X and  $I_{E.coli}^{all}$  is the total number of interactions in *E. coli*. The log2 of the ratio ensures that selection for and against are represented symmetrically in the graph. For example, if  $R_{motif} = 0.9$  and  $R_{all} = 0.6$ , C.I. can be calculated as  $\log_2 (.9/.6) = 0.58$ . Thus if interactions in motifs are selected for, then the C.I. value will be greater than 0, if not, the value will be less than 0. Please refer to appendix A for a detailed procedure.

(a) Interactions in motifs are selected for

- 01: note the number of interactions in motifs in *E. coli*  $I^{em}$  and all interactions in *E. coli*  $I^{ea}$
- 02: for each genome x of the 175 genomes, note the number of interactions conserved,  $I^{xa}$  & interactions conserved that forms a motif in *E. coli*,  $I^{xm}$ .
- 03: if ( $I^{xa} < I^{em}$ )  $R_{motif} = I^{xa}/I^{em}$
- 04: else  $R_{motif} = 1$
- 05:  $R_{all} = I^{xa}/I^{ea}$
- 06: C.I. =  $\log_2 (R_{motif}/R_{all})$
- 07: plot % genes conserved v/s C.I. for every genome

(b) Interactions in motifs are neutrally removed

- 01: note the number of interactions in motifs in *E. coli*  $I^{em}$  and all interactions in *E. coli*  $I^{ea}$
- 02: for each of 10,000 times
- 03: create 175 random networks of similar sizes (one for each of the 175 genomes)
- 04: for each of the 175 networks, note the number of interactions conserved,  $I^{xa}$  & interactions conserved that forms a motif in *E. coli*,  $I^{xm}$ .
- 05:  $R_{motif} = I^{xm}/I^{em}$
- 06:  $R_{all} = I^{xa}/I^{ea}$
- 07: C.I. =  $\log_2 (R_{motif}/R_{all})$
- 08: calculate mean and  $\sigma$  of C.I. for each of the 175 networks over the 10,000 runs
- 08: calculate P-value as the fraction of the runs where C.I. was  $\geq$  observed
- 10: calculate Z-score as  $Z = (\text{observed} - \text{mean})/\sigma$
- 11: plot % genes conserved and the mean C.I. value

(c) Interactions in motifs are selected against

- 01: note the number of interactions in motifs in *E. coli*  $I^{em}$  and all interactions in *E. coli*  $I^{ea}$

- 02:**  $d = I^{ea} - I^{em}$
- 03:** for each of the 175 genomes,  $x$ , note the number of interactions conserved,  $I^{xa}$  & interactions conserved that forms a motif in *E. coli*,  $I^{xm}$ .
- 04:** if  $(I^{xm} > d)$   $R_{\text{motif}} = (I^{xm} - d)/I^{em}$
- 05:** else  $R_{\text{motif}} = 0$
- 06:**  $R_{\text{all}} = I^{xa}/I^{ea}$
- 07:** C.I. =  $\log_2 (R_{\text{motif}}/R_{\text{all}})$
- 08:** plot % genes conserved v/s C.I for every genome

## M10: Procedure to evaluate significance of LSI

1. For each genome studied, a random network was generated using the procedure described in M7. Interaction conservation profile and motif conservation profile were calculated as in M5 and M8. Distance between the organisms was calculated as the Euclidean distance between the interaction (or motif) profile of the two organisms. This gave a distance matrix which was converted into a similarity matrix by normalizing the distance by the maximum distance and then subtracting the value from 1. Organisms were grouped into lifestyle classes and average similarity between organisms in the same lifestyle class and belonging to different lifestyle classes was calculated. This gave rise to the lifestyle similarity matrix. LSI was then calculated as the ratio of the average similarity between organisms of the same lifestyle to the average similarity between organisms belonging to different lifestyles.

Schematically:

		LS1		LS2	
		O1	O2	O3	O4
LS1	O1	1.0	0.8	0.4	0.3
	O2	0.8	1.0	0.5	0.4
LS2	O3	0.4	0.5	1.0	0.7
	O4	0.3	0.4	0.7	1.0

Each element in the matrix represents the normalized similarity between interaction or motif profiles for a given pair of organisms

		LS1	LS2
LS1		0.9	0.4
LS2		0.4	0.85

Average similarity between organisms having different lifestyles

$$LSI = \frac{\text{Average similarity between organisms belonging to the same lifestyle}}{\text{Average similarity between organisms belonging to different lifestyles}} = \frac{\frac{\sum \text{Diagonal elements}}{\text{Number of diagonal elements}}}{\frac{\sum \text{Off diagonal elements}}{\text{Number of off diagonal elements}}}$$

2. The above procedure was done for 1000 times and the p-value was calculated as: Ratio of the number of time LSI exceeds the observed value to the number of trials performed (1000 in this case). Z-score was calculated as:

$$Z\text{-score} = \frac{(\text{observed LSI} - \text{average LSI for the simulation})}{\text{Standard Deviation}}$$

## REFERENCES

Lespinet, O., Wolf, Y. I., Koonin, E. V., and Aravind, L. (2002). The role of lineage-specific gene family expansion in the evolution of eukaryotes. *Genome Res* 12, 1048-1059.

Salgado, H., Gama-Castro, S., Martinez-Antonio, A., Diaz-Peredo, E., Sanchez-Solano, F., Peralta-Gil, M., Garcia-Alonso, D., Jimenez-Jacinto, V., Santos-Zavaleta, A., Bonavides-Martinez, C., and Collado-Vides, J. (2004). RegulonDB (version 4.0): transcriptional regulation, operon organization and growth conditions in *Escherichia coli* K-12. *Nucleic Acids Res* 32, D303-306.

Shen-Orr, S. S., Milo, R., Mangan, S., and Alon, U. (2002). Network motifs in the transcriptional regulation network of *Escherichia coli*. *Nat Genet* 31, 64-68.