A global transcriptional regulatory network for *Escherichia coli* robustly connects gene expression to transcription factor activities

Fang et al.

Supporting Information (SI)

SI Methods.

Transcriptional regulatory network expansion. We constructed an expanded transcriptional regulatory network (TRN) based on RegulonDB version 9.4 (last updated 05-08-2017) (1) and primary literature that included Chromatin ImmunoPrecipitation (ChIP)-binding data. Specifically, we added ChIP-based regulatory interactions for 15 regulons: *arcA* and *fnr* (2–4), *argR* (5, 6), *trpR, lrp* (6), *fur* (7), *gadEWX* (8), *oxyR, soxRS* (9), *purR* (10), *crp* (11) and *cra* (12). All regulatory interactions were specified to be either activation or repression. If the regulatory direction was uncertain, we added an interaction each for activation and repression.

Preparation of expression compendium. We used the EcoMAC microarray compendium (13) to analyze transcriptomic shifts across conditions. In this study, we aimed to assess how consistent our TRN was with measured transcriptome changes across a variety of conditions. Therefore, we chose to exclude experiments from the compendium that either perturbed the TRN wiring, or included artificial environmental perturbations that may not be representative of the evolutionary history of *E. coli*. We also focused our analysis on the exponential growth phase. We thus included only a relevant subset of all conditions, as with (14). Specifically, we excluded regulatory rewiring samples, as they would not represent the naturally-evolved expression patterns. We also removed microgravity and magnetic treatment conditions, as these perturbations were not representative of the evolutionary history of *E. coli*. Since our TRN was reconstructed primarily for *E. coli* K-12, we kept only strains labeled as K12, MG1655, BW25113, and W3110. We removed time-dependent samples (i.e., kept arrays with Time labeled blank, WT, exponential, mid log phase, and mid-log phase). Finally, we had 444 relevant samples.

PCA analysis. PCA analysis was performed with the PCA function in the sklearn.decomposition package. Each principal component is a linear combination of all genes, and 100 genes with the heaviest loadings in each component were subjected to enrichment analysis with respect to regulon, GO, COG and KEGG.

Analysis of the first two principal components from PCA. PCA reduced the dimensionality of the dataset to 50 principal components by 441 samples, in which each principal component was a linear combination of 4,189 genes. Upon plotting the first and second components, the 441 samples were separated into 3 distinct groups (Fig. S18). To understand the separation of the data, metadata including medium, oxygen level, and carbon source were used to label the data points. However, the division of the dataset between the 3 groups did not show a clear correlation with metadata. Interestingly, out of all 441 samples, all 188 samples originating from the Faith lab fell into the same group (Fig. S19). But also note in most of Faith et al.'s samples, DNA damage was induced by norfloxacin, which was not used by any other group. The usage of norfloxacin could potentially explain the clustering of the 188 experiments.

Regulon enrichment analysis was performed on the top 100 genes that carried the most weight in the first and second principal components. The results showed that the first component was enriched for only 3 regulon *nanR* ($P = 0.02$), *basR* ($P = 0.035$), *mlrA* ($P = 0.02$). The second component was enriched for multiple stress response regulons including acid resistance regulons *GadE, GadW*, *GadX* and *phoP*, antibiotic resistance regulon *marA* ($P < 10^{-4}$), anaerobic growth regulon *adiY*, motility system regulons *flhD/flhC and fliZ*. However, the coverage of the transcriptional regulatory network was relatively low for the top loaded genes. Out of the top 100 genes for each component, 69 genes were not found in the TRN for component 1, while 43 genes were not found for component 2.

Non-negative Matrix Factorization (NMF). NMF decomposes the non-negative matrix A into two positive matrices W and H: $A = WH$. Matrix A is generated from the EcoMAC dataset as follows: to meet the non-negativity constraint for NMF analysis, each gene was represented in two columns. One indicates positive expression and the other one indicates negative expression compared to wild type (15). NMF was then performed on the transformed dataset A that had a dimension of (8378, 441). The reduced dimensionality was determined by two methods. The first method compares NMF with singular value decomposition on a random matrix that has the same dimension, mean and variance as the original dataset. The second approach adopted from Wu's study ensures the stability of NMF results (16). Wu's approach minimizes the dissimilarities between matrices across different runs. NMF analysis was then performed using the NMF function in the sklearn decomposition package, with the number of components set to be 40, and initializing method set to be 'nndsvd', which is better for sparseness. Default values were used for all other parameters. To reconstruct matrix W, the negative expression is subtracted from the positive expression for each gene to create a new matrix W that has a dimension of (4189,40). Each column in the W represents a metagene, and the entries represent the coefficient of each gene. Matrix H is the expression pattern of metagenes. The top genes that account for 15% of the weight for the entire metagene were identified as dominating genes, and enriched for regulons.

Selection of dimensionality for NMF. The first method used for dimensionality selection was adopted from Kim and Tidor's paper (17). It utilized singular value decomposition (SVD), one of the more established methods for dimensionality reduction, as SVD was proven to produce the minimum error for a given

dimensionality. It is illustrated in Fig. S20 that NMF is an appropriate method for dimensionality reduction on this dataset, as the error generated during NMF reconstruction is comparable to that produced by SVD. For comparison, SVD is also performed on a random dataset that does not have any correlated features. The random matrix has a gaussian distribution and shares the same dimension, mean and variance of the EcoMAC dataset. The slope of the SVD_random represents the additional structure of the uncorrelated matrix captured by adding one more basis vector. The comparison of the slopes between NMF and SVD_random justified the choices of the reduced dimension, as the slopes are comparable between the dimension of 35 and 50.

Due to the random seed utilized during the Sci-kit Learn NMF function, the decomposition result from each run varies. Wu's method (16) was adopted to ensure the stability of the NMF results. For each dimension between 35 and 50, we generated 100 alternative optimal solutions close to the global optimum. To quantify the stability of the W matrix for each dimension, dissimilarity between W matrices is measured by an Amari-type error, which is calculated from the cross-correlation matrix between the columns of W matrices. The dimension that has the smallest average Amari-type error produces the most stable NMF results. For the EcoMAC dataset, a dimension of 40 produced the smallest Amari-type error and thus was chosen to be the reduced dimension.

Overlap between PC membership and NMF membership. We investigated the overlap of dominant genes between PCA and NMF analysis. We have identified a total of 1961 dominant genes (following the same method in NMF analysis) in the 40 principal components in PCA analysis, and 1734 dominant genes in 40 metagenes in NMF analysis. Our results showed that 80.2% (1391/1734) of the dominant genes in NMF analysis overlapped with the dominant genes in PCA, which suggest consistency between two methods. Due to the difference in the nature of these two methods, the individual components of these two methods are different as expected.

Regulon Enrichment analysis. Regulon enrichment analysis was performed using the fisher_exact function in the scipy.stats package. Prior to the analysis, the following variables were calculated: the size of each regulon, the size of the set of genes subjected to enrichment analysis, the overlap between the set of interest and each regulon, and the total number of genes involved in TRN and EcoMAC. The Fisher-exact test calculates P-values for each regulon, and a regulon is considered to be enriched if the p-value is less than 0.05.

Core regulatory module identification. We identified a core regulatory network by integrating the results of differential gene expression identification, enrichment analysis of ChIP-based regulons, and dimensionality reduction into biological 'parts' by NMF. For NMF, we used regularized NMF with a 15% cutoff of metagene loadings to define representative genes. Alternatively, we used non-smooth NMF (18), which produced more sparse metagenes and required a 0.001 cutoff of the coefficients. Since the NMF algorithms use randomization to solve the non-convex optimization problem, we randomly started NMF 100 times to retain alternate optimal solutions close to the global optimum. We then performed enrichment analysis of regulons for the representative genes in the metagene loadings.

Using all 200 (2 NMF algorithms x 100 runs) TF-metagene enrichment results, we created a co-enrichment network of TFs: i.e., co-occurrence network of pairs of TFs enriched in the same metagene. We quantified the strength of co-enrichment of a given TF pair using the Jaccard index

$$J(A,B) = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}$$

where $|A \cap B|$ is the number of metagenes for which TF A and B are co-enriched, $|A|$ the number of metagenes for which TF A is enriched, and $|B|$ the number of metagenes for which TF B is enriched. To retain only statistically significant pairs of co-enriched TFs, we compared the network against 100,000 randomly generated co-enrichment networks by sampling from the observed frequency of enriched TFs. Only the TF pairs having FDR-adjusted $P < 0.05$ were kept. We also only kept TF pairs that were strongly coenriched, in this case Jaccard index > 0.18. We finally had 522 significant co-enriched TF pairs. We then performed community detection on this significant co-enrichment network. We used multi-level modularity optimization using the cluster_louvain function in igraph (19). The modularity coefficient was 0.483. The modularity for the computed graph was always greater than the random graphs, so we deemed it to be significant. Finally, we identified a significant, core TRN consisting of 10 major modules (11, including non-coenriched TFs) that were functionally-annotated by DAVID(20) followed by manual curation. These modules were then used for further characterization.

We applied this workflow also for the COLOMBOS compendium (21). For COLOMBOS, we obtained 484 significant co-enriched TF pairs. The multi-level modularity optimization resulted in 11 modules (12, including non-coenriched TFs) with modularity coefficient of 0.57. The regulatory modules for both compendia are in Dataset S2.

Robustness of TF modules. Since new ChIP-binding data is constantly being generated, the TRN network is always expanding to incorporate new interactions. Therefore, we also evaluated the robustness of the TRN modules when new interactions between TFs and genes are added to the TRN network. We added in low-confidence interactions from up to 60 random regulons ten times each. Note that if the TF for a randomly chosen regulon already existed, the low-confidence interactions for that regulon were added. We then computed similarity of clusterings using two measures. First, we used the variation of information (VI) (22) for the TFs common between the original and perturbed TRNs. This metric thus reflects to what extent TFs within a module get re-assigned to different modules as new regulatory interactions or TFs are discovered. Because VI did not account for the new TFs added we also used a Jaccard index-based metric to quantify the overall change in TF modules. Given original and alternative modules from the original and perturbed TRNs, we computed the Jaccard index for all pairs of modules based on the TFs within the modules. For each original module, we defined its similarity to be the highest Jaccard index between it and all alternate modules. Thus, the similarity of the original to the perturbed modules was the mean of these Jaccard indices across all the original modules. The similarity of the alternative to original modules was computed in the same way, to account for potentially new modules arising with new TFs. The final Jaccard index-based

similarity between the two clusterings was the average of these two similarities.

Conservation of TF regulons. Gene annotation of strains and species were obtained from the SEED server (http://theseed.org), and ortholog calculation to *E. coli* K-12 MG1655 was also performed on the RAST (Rapid Annotation using Subsystem Technology) server (23). The number of strains in each phylogenetic group was 33 Enterobacteriaceae, 134 γ-protebacteria, 40 β-proteobacteria, 58 α-proteobacteria, and 23 δ-proteobacteria. The percentage of gene conservation indicates the number of strains having a particular gene in a phylogenetic group divided by the total number of strains in the group. We computed conservation of the 147 *E. coli* TFs in our hiTRN across these phylogenetic groups. We then identified modules consisting of TFs having significantly high or low evolutionary conservation compared to all other TFs (Wilcoxon rank sum test $P < 0.05$).

TF binding motif analysis. Sequences of TF binding motifs were collected from RegulonDB(1). Sequence homology between TF binding motifs was analyzed using the global alignment function in the Bio.pairwise2 package. The best match between each pair of TFs was identified and alignment score was recorded. For each TF module, we compared the alignment scores of TF within and outside the modules with Mann-Whitney-Wilcoxon test. TF modules that have a p-value less than 0.05 were considered to have more similar binding motifs within the modules.

TF structure analysis. Homology models for the protein sequences were generated for 117 of the 147 TFs in the core regulatory network using the I-TASSER software package (24). DNA-binding domain predictions were also carried out for 114 TFs based on templates available in the PDB. The structures were compared using the pairwise rigid FATCAT aligner (25), creating an all vs. all alignment. The average TM-score (similarity score in the range of [0,1] (26)) of all pairwise alignments within each cluster was compared to the average TM-score of the alignments from randomly-generated groups of TFs of the same size to generate a p-value for each cluster. In addition, a hypergeometric test was applied to domain assignments by (27) to test for domain enrichment in clusters.

Toxin-Antitoxin analysis on TF modules . In the most updated TRN, we have included multimer TFs, among which 3 of them are also toxin-antitoxin (TA) gene pairs: *dinJ-yafQ, relE-relB, yefM-yoeB*. We also included monomer TFs that are components of the TA pairs, including *yefM, relB, higA*. Out of 6 TFs that are members of TA pairs, 5 TFs (*relE-relB, yefM, relB,yefM-yoeB, higA*) are in the same TF module (module 6). Interestingly, module 6 is represented by stress response TFs, which is consistent with the functional roles of the TA members (amino acid starvation, multidrug resistance, etc.). Thus, the TAs studied here change expression in a functionally cohesive manner across conditions.

Differentially-expressed gene (DEG) Identification. DEGs were identified using the R package limma in Bioconductor (28). The reference samples used for all samples were wild type MG1655 grown in M9 with glucose as carbon source under aerobic condition. The replicates of all experimental conditions were identified and compared against reference sample using limma. Genes having an expression level fold change greater than 2 and (FDR) adjusted p-value less than 0.05 were identified as DEGs.

Network-expression consistency analysis. Network analysis was done on the DEGs for experiments that involve at least one TF knockout. DEGs were identified using limma (28) with different reference samples for each experimental condition. Using SigNetTrainer (29), we computed the consistency of the TRN (including direction of regulation–activation vs. inhibition) with measured expression changes for TF knockout experiments.

In addition, consistency and reachability for only DEGs were calculated for 20 experiments, as 3 experiments had none or only very few DEGs identified. We have also performed a permutation test, in which we selected a random TF (or two TFs depending on the original number of TFs that were knocked out in each experiment) to be the knocked out TF in each experiment, and calculated the reachability from DEGs to the randomly selected TF(s). P-value for the permutation test was calculated for 10,000 runs to be 3.91×10^{-4}.

Reachability was calculated by utilizing the igraph package in Python (19). A graph containing all the nodes and edges in the TRN was established, and all the nodes that could be reached from each TF were recorded. Reachability was then calculated by identifying the overlap of the set of nodes reachable by the TF and the DEGs in the TF knockout experiment.

Information Analysis. The mutual information between two distributions is defined as:

$$MI(X,Y) = H(X) + H(Y) - H(X;Y)$$

where H(X) is the entropy of distribution X, H(Y) is the entropy of distribution Y, and H(X,Y) is the joint entropy of distributions X and Y. The entropy of a discrete distribution is defined as:

$$H(X) = \sum_i -p_i \log p_i$$

where p_i is the probability of state i. Mutual information for continuous variables can be calculated using differential entropy, rather than entropy. The mutual information between two genes was defined as the mutual information between the log fold change expression profiles of each gene, as calculated by the NPEET package for Python (30).

For each TF, the mutual information was calculated between the TF expression profile and the expression profile of each gene in its regulon. This distribution was compared against the MI between the TF expression profile and all other genes not in its regulon using the Wilcoxon rank-sum test ($\alpha = 0.05$). The null hypothesis states the MI distributions originated from the same distribution, and the alternative states that the MI distribution of the genes in the regulon is greater than the distribution of the genes outside the regulon. Only high confidence interactions were included in the analysis.

In addition, the mutual information was compared for pairs of genes in the same regulatory module as compared to genes not sharing a module. The mutual information was calculated for 1,000 randomly selected gene pairs in each module, and for 1,000 randomly selected gene pairs that did not belong to

the same module, serving as the null distribution. A Mann-Whitney-U test was applied to each module against the null, with a significance value of $p < 0.05$ to determine if the MI between genes within each module were significantly higher than the MI between genes not sharing a module.

Expression Profile Regression. The expression log fold change of transcription units was calculated by averaging the log fold change of each gene in the TU. TUs were defined from RegulonDB (1), and only those with strong evidence or greater were kept; in all other cases the TUs were defined as single genes. Of the resulting 1538 TUs, EcoMAC contained expression data for 1364 TUs, and sigma factors were defined for 1098 TUs.

Eight model structures were used to predict the TU expression profile, with features including the log-fold change of the known regulators of the TU, cooperation/competition terms for all combinations of two TFs, and the log-fold change of the known sigma factors of the TUs.

Both linear regression and support vector regression were performed using the Scikit-learn package for Python (31). For the support vector regressors, the parameters C, gamma and epsilon were optimized using 3-fold cross validation for each individual TU regression. The accuracy of the regression models was measured by the average coefficient of determination (R^2) across a 10-fold cross validation. Samples from the same lab under the same condition were not split across folds.

We performed an F-test on the linear regression of the training data to determine if the TFs or sigma factors significantly improved the prediction results for each gene. The R^2 values of the testing data as predicted by the linear model and SVR were compared using the Wilcoxon signed-rank test. To determine whether the model captured condition-specific effects, we shuffled the TU expression profiles 1000 times and then ran the regression on each shuffled profile using 10-fold cross validation, while maintaining the condition-based order of the regulator expression profiles. The shuffling served to unlink the experimental condition of the regulators from the conditions of the predicted expression profile. Significance was assigned to each TU by calculating the fraction of shuffled profiles with a higher testing R^2 value than the original regression, and applying the Benjamini-Hochberg procedure to the resulting distribution with an FDR of 0.05. The relative power of our TRN compared to a randomized TRN was calculated by randomly assigning 1000 sets of TFs to each TU and running the regression using both the linear model and the SVR on each set. The number of regulators for each TU was maintained, and TFs that had a high mutual information with true regulators of the TU were not assigned to the TU. As before, significance was assigned to each TU by determining the fraction of randomly generated TRNs with higher testing R^2 values than the original regression using a Benjamini-Hochberg procedure with an FDR of 0.05.

Regression Model Selection. We implemented eight regression models to predict gene expression profiles from the EcoMAC dataset. All eight models, four linear regressors, two SVRs with linear kernels, and two SVRs with gaussian kernels, predicted gene expression profiles from the gene's TF expression profiles. Four models included the gene's known sigma factors (32), and two of the linear models accounted interactions between TFs as shown below:

Model 1: Linear Model

$$Y_i = a_i + \sum_{j=1}^{n} b_{ij} y_{\mathrm{TF}j}, \quad [1]$$

Model 2: Linear Model with Sigma Factors

$$Y_i = a_i + \sum_{j=1}^{n} b_{ij} y_{\mathrm{TF}j} + \sum_{j=1}^{m} c_{ij} y_{\sigma j}, \quad [2]$$

Model 3: Linear Model with TF Interaction

$$Y_i = a_i + \sum_{j=1}^{n} b_{ij} y_{\mathrm{TF}j} + \sum_{j=1}^{n} \sum_{jk=1}^{n} d_{ijk} y_{\mathrm{TF}j} y_{\mathrm{TF}k}, \quad [3]$$

Model 4: Linear Model with TF Interaction and Sigma Factors

$$Y_i = a_i + \sum_{j=1}^{n} b_{ij} y_{\mathrm{TF}j} + \sum_{j=1}^{m} c_{ij} y_{\sigma j} + \sum_{j=1}^{n} \sum_{k=1}^{n} d_{ijk} y_{\mathrm{TF}j} y_{\mathrm{TF}k}, \quad [4]$$

Model 5: Linear SVR

$$Y_i = f(y_{\mathrm{TF}1}, y_{\mathrm{TF}2}, \dots), \quad [5]$$

Model 6: Linear SVR and Sigma Factors

$$Y_i = f(y_{\mathrm{TF}1}, y_{\mathrm{TF}2}, \dots, y_{\sigma 1}, \dots), \quad [6]$$

Model 7: SVR with Gaussian Kernel

$$Y_i = f(kernel(y_{\mathrm{TF}1}, y_{\mathrm{TF}2}, \dots)), \quad [7]$$

Model 8: SVR with Gaussian Kernel and Sigma Factors

$$Y_i = f(kernel(y_{\mathrm{TF}1}, y_{\mathrm{TF}2}, \dots, y_{\sigma 1}, \dots)), \quad [8]$$

where Y_i is the expression profile of gene i, $y_{\mathrm{TF}j}$ is the expression profile of TF_j, $y_{\sigma j}$ is the expression profile of sigma factor j, a_i is the baseline expression level for gene i, b_{ij} is the coefficient of TF_j on gene i, c_{ij} is the coefficient of sigma factor j on gene i, d_{ijk} is the interaction term between gene i, TF_j, and TF_k, and *kernel* is the gaussian kernel transformation. The TF interaction terms were not required for the SVRs as a gaussian kernel can account for nonlinearities and interplay between regressors.

The models were evaluated using 10-fold cross validation, with samples from the same lab under the same conditions grouped in the same fold. We performed an F-test of overall significance on Model 3 (Linear Model with TF Interaction) to determine whether the model fit the data better than an intercept-only model. In addition, an F-test of overall significance was applied to an additional sigma factor-only linear model (with interactions) to highlight the effects of including sigma factors as regressors. We then compared the linear model with the best accuracy (Model 4) to the SVR with the best accuracy (Model 8) to compare the strength of each algorithm using the Wilcoxon rank-sum test. The model with the highest overall accuracy on the testing dataset (Model 8) was used for the remainder of the analysis.

When shuffling the conditions, the TU expression profile was shuffled 1000 times, while keeping the same TRN structure. P-values were determined by the number of these trials that resulted in a higher R^2 value than the original model, and significance was assigned based on the Benjamini-Hochberg

procedure with an FDR of 0.05. To compare our TRN against a randomly generated TRN, a pool of all transcription factors were generated for each TU excluding those known to regulate the TU, any TFs in the TU, and any TFs that had a high mutual information with a TF known to regulate the TU. A pair of TFs with mutual information above the 75th percentile was designated as a high mutual information pair (see Fig. S13). Five hundred sets of transcription factors were selected from this pool, all with the same number of regulators as defined in the original TRN. As before, significance was assigned to each TU by determining the fraction of randomly generated TRNs with higher testing R^2 values than the original regression using a Benjamini-Hochberg procedure with an FDR of 0.05.

Surrogate Variable Analysis. Surrogate variable analysis was performed as described in Leek and Storey (33). The residuals for the analysis were generated from the SVR model with sigma factors. Three surrogate variables were identified and compared against the compendium metadata.

Comparing hiTRN with only high-confidence interactions in RegulonDB. To further characterize the additional information added by high-confidence interactions identified by ChIP data, we performed the same analysis on EcoMAC with a new TRN that only contains the high-confidence interactions in RegulonDB. Since the input for network analysis is different (hiTRN versus RegulonDB network) for sigNetTrainer, it is difficult to compare the results for Figure 2. Instead, we made a table that compares the number of differentially-expressed genes that can be reached from the knocked-out TF in these two networks (see SI Table 1). Results suggested that numbers of DEGs decreased in most experiments when excluding ChIP-based interaction, especially experiments involving *arcA* and *fnr*.

Lastly, we did regression analysis on EcoMAC using the TRN with only RegulonDB interactions (Fig. S10). 596 of the 690 TUs with known regressors (86%) yielded significant differences between the shuffled expression profile regression and the original regression (FDR-adjusted $P < 0.05$), and 90 TUs (13% of 690) were predicted significantly better than random TRNs for the best SVR (FDR-adjusted $P < 0.05$), which is similar to the results from the hiTRN.

Comparison with COLOMBOS dataset. To validate our results, we performed the same analysis on a different *E.coli* expression dataset COLOMBOS (21). We have used the same filtering standard as EcoMAC and calculated the missing data with the R package Impute (34). The processed COLOMBOS dataset has 4266 genes and 2049 profiles. We first selected the dimension of NMF reduction following Kim and Tidor's method (17) to be 63, as many additional conditions are incorporated in COLOMBOS.

Regulatory modules: After reducing the COLOMOBOS dataset using NMF, we ran the enrichment analysis on dominant genes of metagenes followed by community detection to identify TF modules. We compared the regulatory modules identified from EcoMAC and COLOMBOS (SI Dataset XX) using the variation of information (VI) (22). VI is a widely used metric to compare how similar two clusterings are:

$$VI(X;Y) = H(X) + H(Y) - 2I(X,Y)$$

For n elements to cluster (i.e., genes), VI is bounded by $\ln(n)$. Alternatively, for k maximum clusters, VI is bounded by $2\ln(k)$. A normalized VI relative to either n or k is bounded between 0 and 1, where 0 indicates equivalent clusterings and 1 indicates zero mutual information between the two clusterings. We computed VI using the clusters.stats function from the fpc R library (35), excluding the unclustered "noise class" from computations. There were $n = 3070$ genes that were common between the regulons contained in both regulatory module sets, from the two expression compendia. The VI normalized by n genes was 0.17. Alternatively, the VI normalized by the number of modules ($k = 11$) was 0.29. Both normalized VI values were significantly lower than 10,000 randomly generated regulatory modules (permutation test, $P < 10^{-4}$). We generated random networks preserving the number of nodes and edges but with randomly re-assigned edges and edge weights randomly sampled within the range of observed weights. We then computed the VI between these random modules and the EcoMAC-based core TRN and compare the VI against that between EcoMAC and COLOMBOS. Based on these tests, we finally concluded that the core TRN identified was significantly preserved regardless of the transcriptomics data set used.

TRN coverage: In addition, we have also extracted 9 TF knockout experiments from COLOMBOS dataset. Excluding the experiments with missing reference sample and no DEG identified, we calculated the consistency and reachability of DEGs in hiTRN to knocked out TF for 4 experiments (see Fig. S2) using SigNetTrainer (29). The result is similar to the 21 TF knockout experiments we previously analyzed. Consistency between prediction and experimental measurement is between 59% and 99%, while 56% of DEGs can be traced back to the knocked out TF (only considering 3 experiments that have more than 10 DEGs identified).

Quantitative gene expression prediction: Moreover, we also performed the regression analysis on this dataset S9. 1081 of the 1375 TUs with known regressors (79%) yielded significant differences between the shuffled expression profile regression and the original regression (FDR-adjusted $P < 0.05$), and 122 TUs (9% of 1375) were predicted significantly better than random TRNs for the best SVR (FDR-adjusted $P < 0.05$). Using only strong interactions from RegulonDB S11, 553 of the 690 TUs with known regressors (80%) yielded significant differences between the shuffled expression profile regression and the original regression, and 85 TUs (12% of 690) were predicted significantly better than random TRNs. These statistics are close to the values generated from the EcoMAC dataset. The mutual information analysis showed that 26% (36/137) of known TFs shared significantly higher MI with genes inside as compared to outside their regulons (FDR < 0.05), which is also on par with the data generated from EcoMAC (28% or 39/137).

Comparing TF modules with previous works. We compared the TF modules we have identified with previous work done by other groups. Baliga lab has identified 590 conditionally co-regulated modules (corems) from a gene expression compendium (36). We compared the corems with the TF modules we proposed by performing enrichment analysis of TF modules for each corem on the gene level. The results showed that 230/590 corems are enriched by at least 1 TF module (see Dataset S3), which shows correlation between some corems and TF modules. For the rest of the corems that are not enriched for any TF modules, potential explanations are: 1. Since we only included high-confidence interactions in hiTRN, not all the

genes in corems are part of the hiTRN; 2. The *E.coli* expression compendium used to create corems has more conditions(e.g. heat, pH, metal) than EcoMAC, so it is possible that the proposed TF modules did not incorporate information related to such conditions. Thus, using a larger compendium with more conditions to create the TF modules could potentially improve the results.

SI Appendix References

1. Gama-Castro S, et al. (2015) Regulondb version 9.0: high-level integration of gene regulation, coexpression, motif clustering and beyond. *Nucleic Acids Research* p. gkv1156.
2. Federowicz S, et al. (2014) Determining the control circuitry of redox metabolism at the genome-scale. *PLoS Genet.* 10(4):e1004264.
3. Myers KS, et al. (2013) Genome-scale analysis of escherichia coli FNR reveals complex features of transcription factor binding. *PLoS Genet.* 9(6):e1003565.
4. Park DM, Akhtar MS, Ansari AZ, Landick R, Kiley PJ (2013) The bacterial response regulator ArcA uses a diverse binding site architecture to regulate carbon oxidation globally. *PLoS Genet.* 9(10):e1003839.
5. Cho S, et al. (2015) The architecture of ArgR-DNA complexes at the genome-scale in escherichia coli. *Nucleic Acids Res.* 43(6):3079–3088.
6. Cho BK, Federowicz S, Park YS, Zengler K, Palsson BO (2012) Deciphering the transcriptional regulatory logic of amino acid metabolism. *Nat. Chem. Biol.* 8(1):65–71.
7. Seo SW, et al. (2014) Deciphering fur transcriptional regulatory network highlights its complex role beyond iron metabolism in escherichia coli. *Nat. Commun.* 5:4910.
8. Seo SW, Kim D, O'Brien EJ, Szubin R, Palsson BO (2015) Decoding genome-wide GadEWX-transcriptional regulatory networks reveals multifaceted cellular responses to acid stress in escherichia coli. *Nat. Commun.* 6:7970.
9. Seo SW, Kim D, Szubin R, Palsson BO (2015) Genome-wide reconstruction of OxyR and SoxRS transcriptional regulatory networks under oxidative stress in escherichia coli K-12 MG1655. *Cell Rep.* 12(8):1289–1299.
10. Cho BK, et al. (2011) The PurR regulon in escherichia coli K-12 MG1655. *Nucleic Acids Res.* 39(15):6456–6464.
11. Latif H, et al. (year?) "chip-exo interrogation of crp, dna, and rnap holoenzyme interactions". *bioRxiv* doi:10.1101/069021.
12. Kim D, et al. (2016) Systems assessment of transcriptional regulation on central carbon metabolism by cra and CRP. *bioRxiv* doi:10.1101/080929.
13. Carrera J, et al. (2014) An integrative, multi-scale, genome-wide model reveals the phenotypic landscape of escherichia coli. *Mol. Syst. Biol.* 10(7):735.
14. Yang L, et al. (2015) Systems biology definition of the core proteome of metabolism and expression is consistent with high-throughput data. *Proceedings of the National Academy of Sciences* 112(34):10810–10815.
15. Kim PM, Tidor B (2003) Subsystem identification through dimensionality reduction of large-scale gene expression data. *Genome Res.* 13(7):1706–1718.
16. Wu S, et al. (2016) Stability-driven nonnegative matrix factorization to interpret spatial gene expression and build local gene networks. *Proc. Natl. Acad. Sci. U. S. A.* 113(16):4290–4295.
17. Kim PM, Tidor B (2003) Subsystem identification through dimensionality reduction of large-scale gene expression data. *Genome Res.* 13(7):1706–1718.
18. Pascual-Montano A, Carazo JM, Kochi K, Lehmann D, Pascual-Marqui RD (2006) Nonsmooth nonnegative matrix factorization (nsNMF). *IEEE Trans. Pattern Anal. Mach. Intell.* 28(3):403–415.
19. Csardi G, Nepusz T (2006) The igraph software package for complex network research. *InterJournal, Complex Systems* 1695(5):1–9.
20. Dennis G, et al. (2003) David: database for annotation, visualization, and integrated discovery. *Genome Biology* 4(9):R60.
21. Moretto M, et al. (2016) Colombos v3. 0: leveraging gene expression compendia for cross-species analyses. *Nucleic acids research* 44(D1):D620–D623.
22. Meilă M (2003) Comparing clusterings by the variation of information in *Learning theory and kernel machines*. (Springer), pp. 173–187.
23. Aziz RK, et al. (2008) The rast server: rapid annotations using subsystems technology. *BMC genomics* 9(1):75.
24. Roy A, Kucukural A, Zhang Y (2010) I-TASSER: a unified platform for automated protein structure and function prediction. *Nat. Protoc.* 5(4):725–738.
25. Ye Y, Godzik A (2003) Flexible structure alignment by chaining aligned fragment pairs allowing twists. *Bioinformatics* 19 Suppl 2:ii246–55.
26. Xu J, Zhang Y (2010) How significant is a protein structure similarity with TM-score = 0.5? *Bioinformatics* 26(7):889–895.
27. Madan Babu M, Teichmann SA (2003) Functional determinants of transcription factors in escherichia coli: protein families and binding sites. *Trends Genet.* 19(2):75–79.
28. Ritchie ME, et al. (2015) limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* 43(7):e47.
29. Melas IN, Samaga R, Alexopoulos LG, Klamt S (2013) Detecting and removing inconsistencies between experimental data and signaling network topologies using integer linear programming on interaction graphs. *PLoS Comput. Biol.* 9(9):e1003204.
30. Kraskov A, Stögbauer H, Grassberger P (2004) Estimating mutual information. *Physical Review E* 69(6):066138.
31. Pedregosa F, et al. (2011) Scikit-learn: Machine learning in python. *J. Mach. Learn. Res.* 12(Oct):2825–2830.
32. Cho BK, Kim D, Knight EM, Zengler K, Palsson BO (2014) Genome-scale reconstruction of the sigma factor network in escherichia coli: topology and functional states. *BMC Biol.* 12:4.
33. Leek JT, Storey JD (2007) Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet.* 3(9):1724–1735.
34. Hastie T, Tibshirani R, Narasimhan B, Chu G (2012) Impute: imputation for microarray data. *R package version* 1(0).
35. Hennig C (2015) *fpc: Flexible Procedures for Clustering*. R package version 2.1-10.
36. Brooks AN, et al. (2014) A system-level model for the microbial regulatory genome. *Mol Syst Biol* 10(7):740.

SI Figures

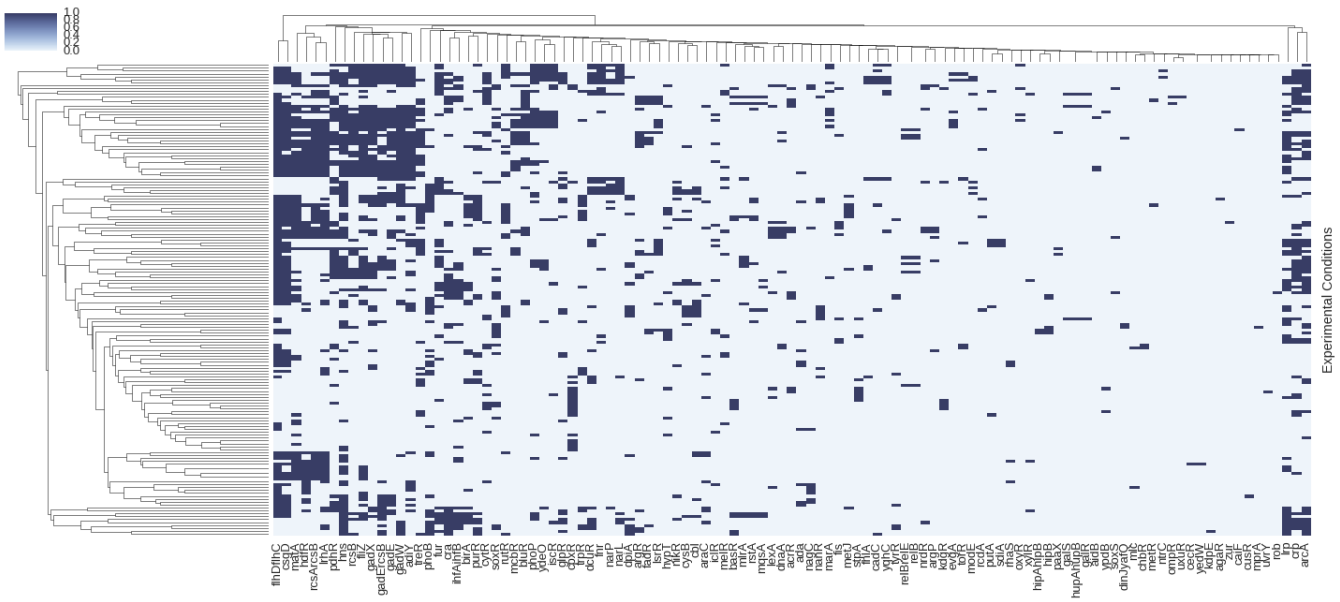

Fig. S1. Regulon enrichment of DEGs under various experimental conditions. Samples that do not have regulons enriched and regulons that are not enriched in any samples are not shown in this graph. Dark and light blue represent enrichment and no enrichment respectively.

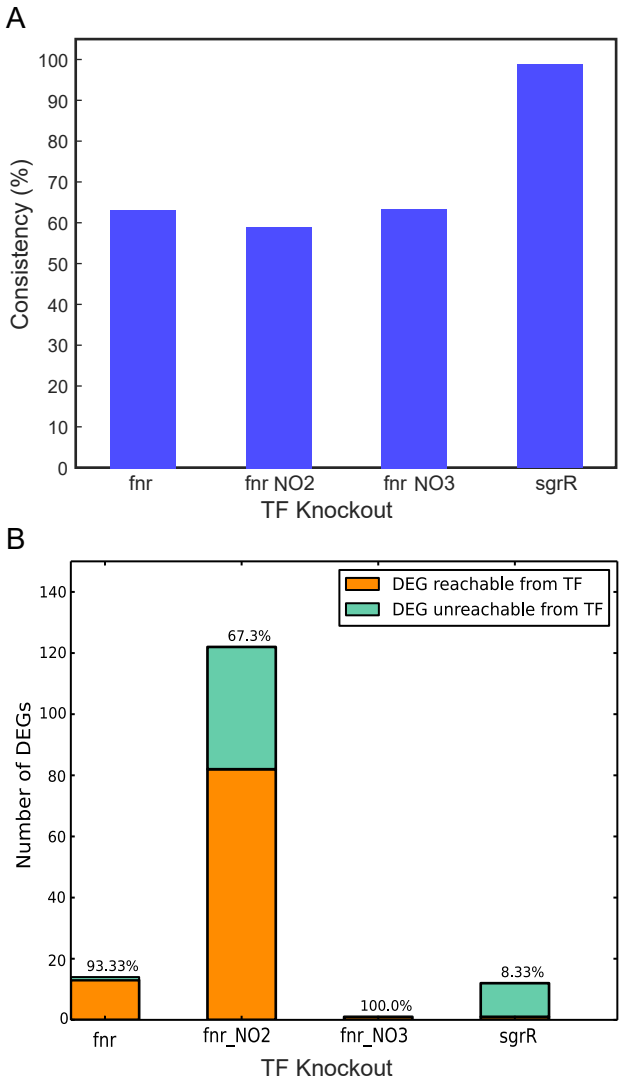

Fig. S2. Consistency of hiTRN with observed differential gene expression in TF knockout experiments from COLOMBOS dataset. (A) Consistency of hiTRN with observed differential and non-differential gene expression accounting for regulatory bias (sign consistency). (B) Reachability from deleted TFs to DEGs in the TRN. Percentages above each bar are the percentage of DEGs that were reachable from the deleted TF within the TRN.

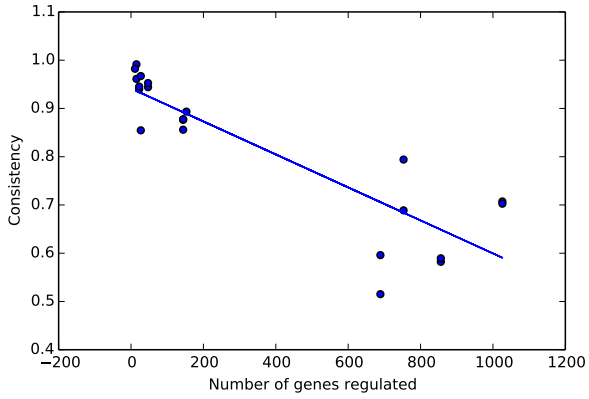

Fig. S3. Correlation between the number of genes regulated and consistency between TRN prediction and experimental measurement. The Pearson correlation coefficient is -0.875 and the p-value is 2.10×10^{-7}.

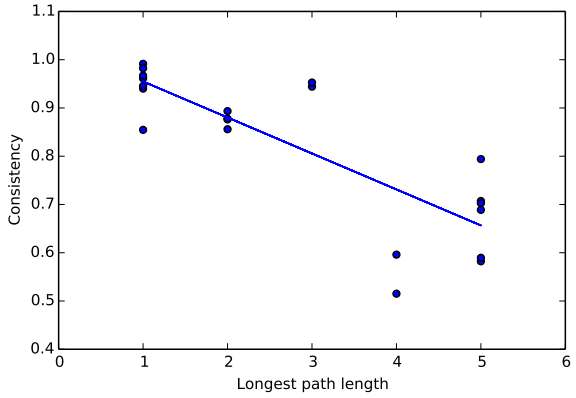

Fig. S4. Correlation between the length of the longest regulatory path of a TF and sign consistency between TRN prediction and experimental measurement. The Pearson correlation coefficient is -0.82 and the p-value is 5.07×10^{-6}.

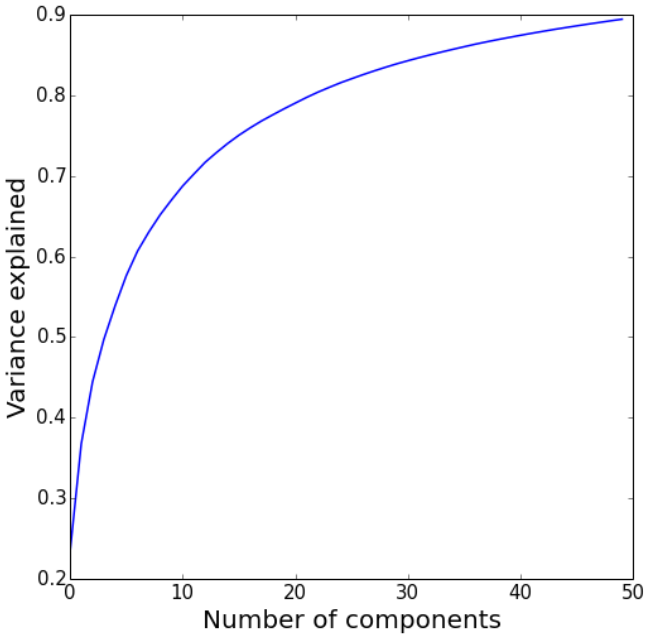

Fig. S5. Variance explained by principal components. The first 40 principal components explain 88% of the variance of the EcoMAC dataset.

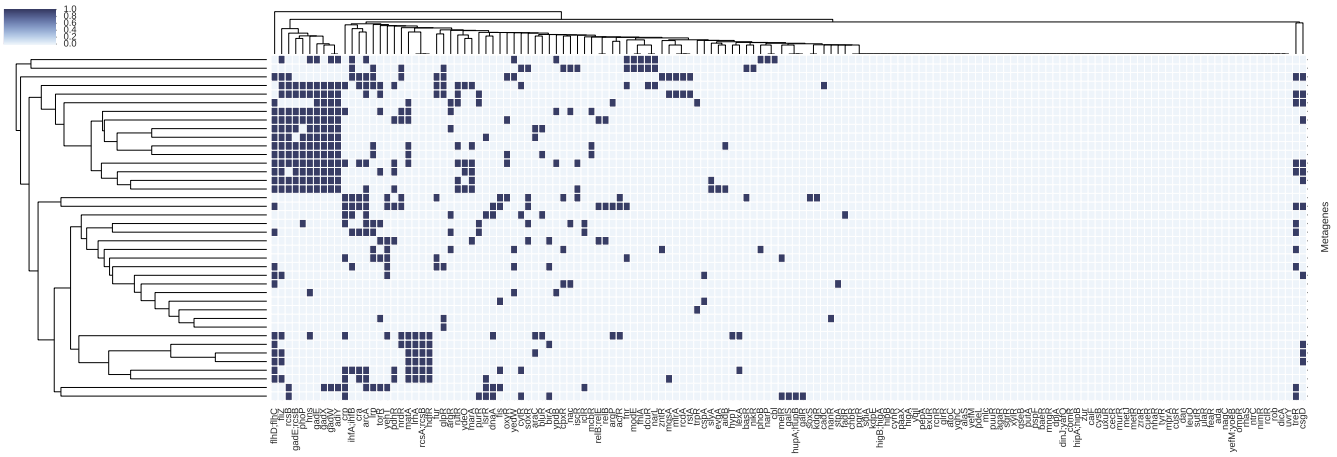

Fig. S6. Regulon enrichment on 40 metagenes identified from EcoMAC. Enrichment on all 147 regulons are shown here.

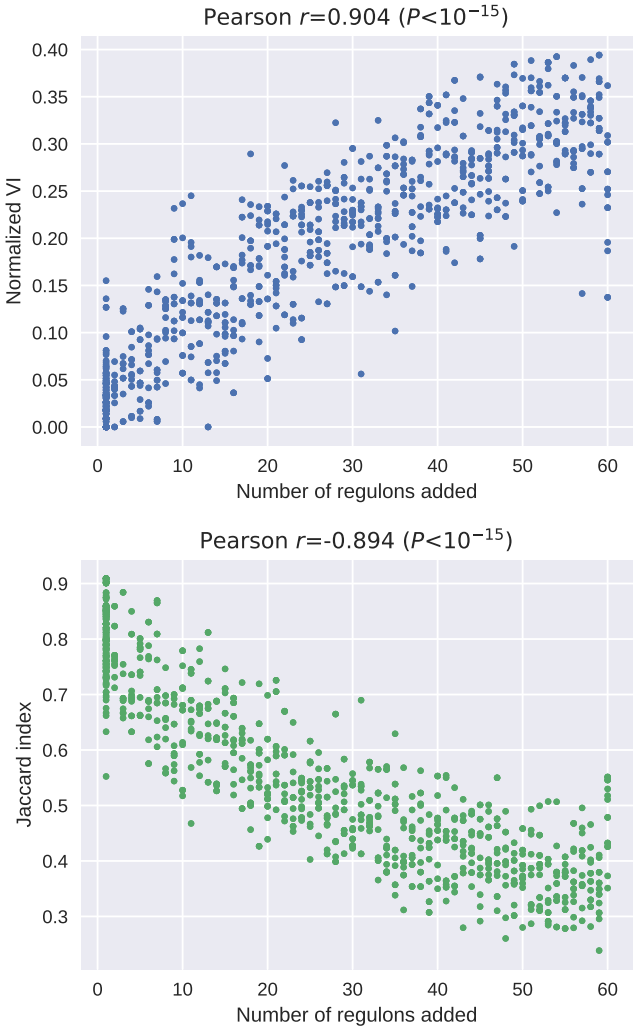

Fig. S7. Stability of TF modules as quantified by normalized variation of information and Jaccard index of when low-confidence interactions were added to the hiTRN from up to 60 random regulons, 10 times each.

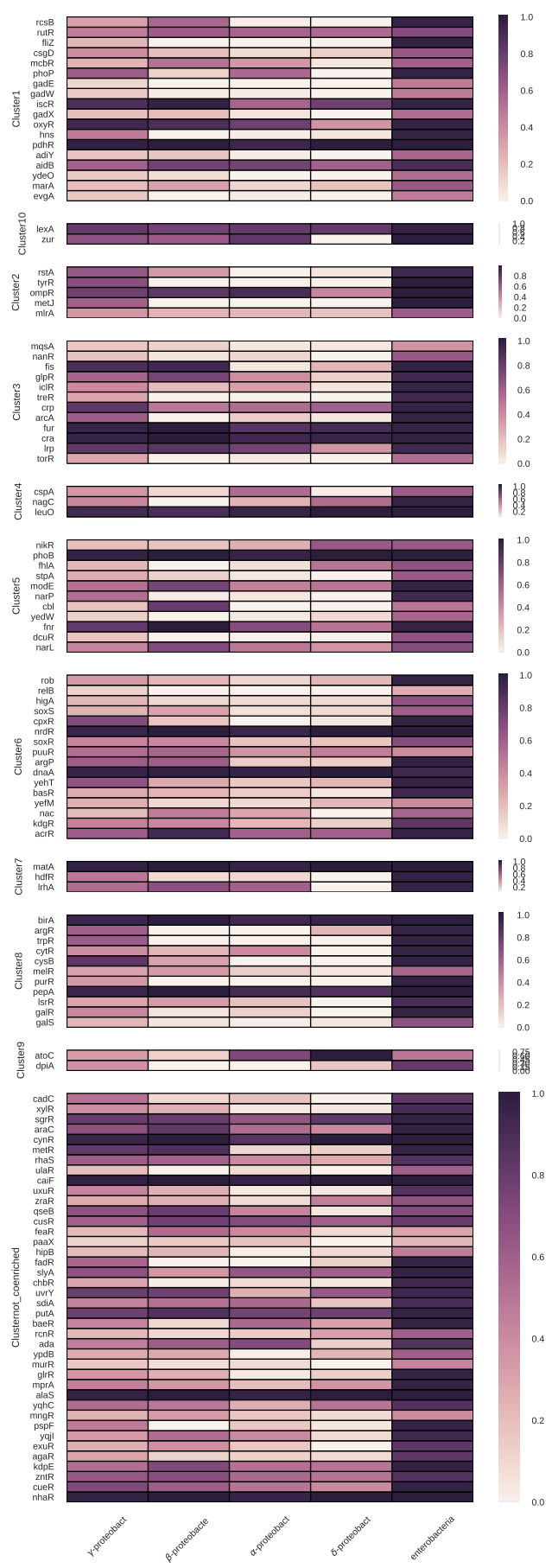

Fig. S8. Evolutionary conservation of TF clusters.

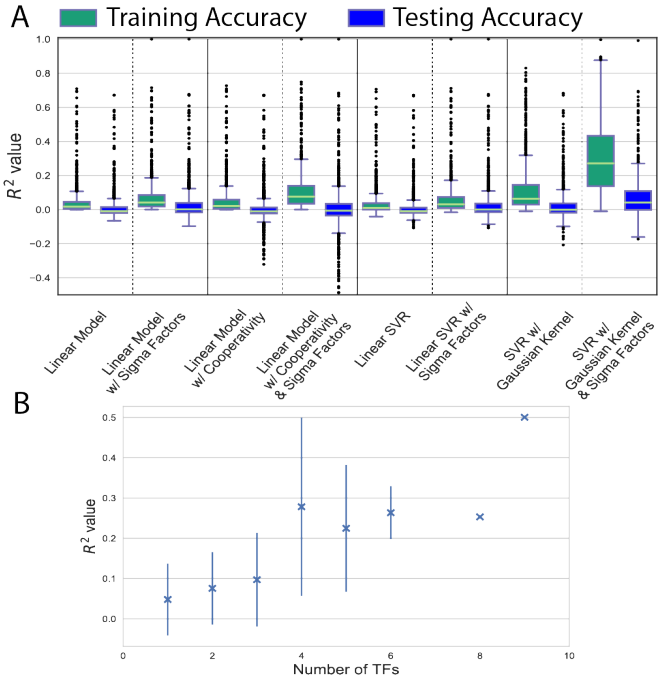

Fig. S9. Accuracy of expression predictions on training and held-out testing transcription units. (A) R^2 of predicted expression profile vs. true expression profile using various regression models from COLOMBOS dataset. (B) R^2 value of the testing dataset predicted by a gaussian kernel SVR, grouped by number of known TFs. Error bars indicate standard deviation for groups with >3 observations.

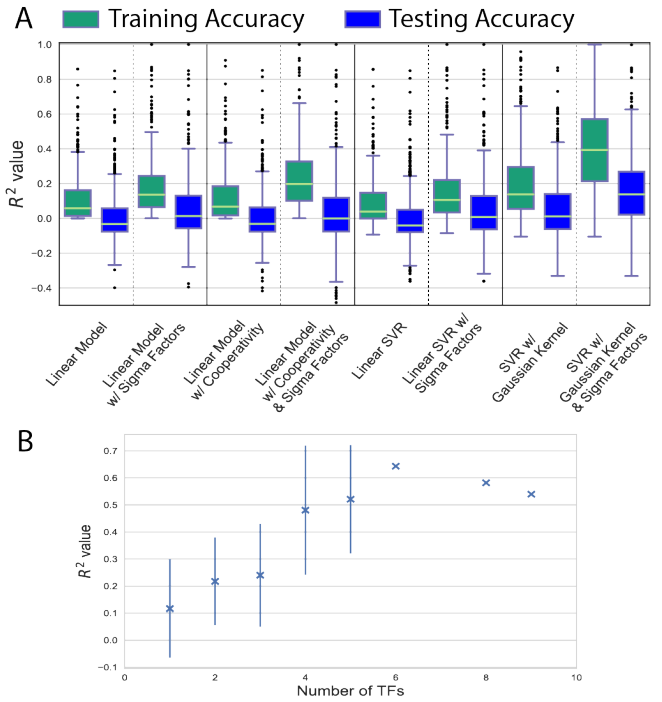

Fig. S10. Accuracy of expression predictions on training and held-out testing transcription units using only strong interactions from RegulonDB. (A) R^2 of predicted expression profile vs. true expression profile using various regression models from EcoMAC dataset. (B) R^2 value of the testing dataset predicted by a gaussian kernel SVR, grouped by number of known TFs. Error bars indicate standard deviation for groups with >3 observations.

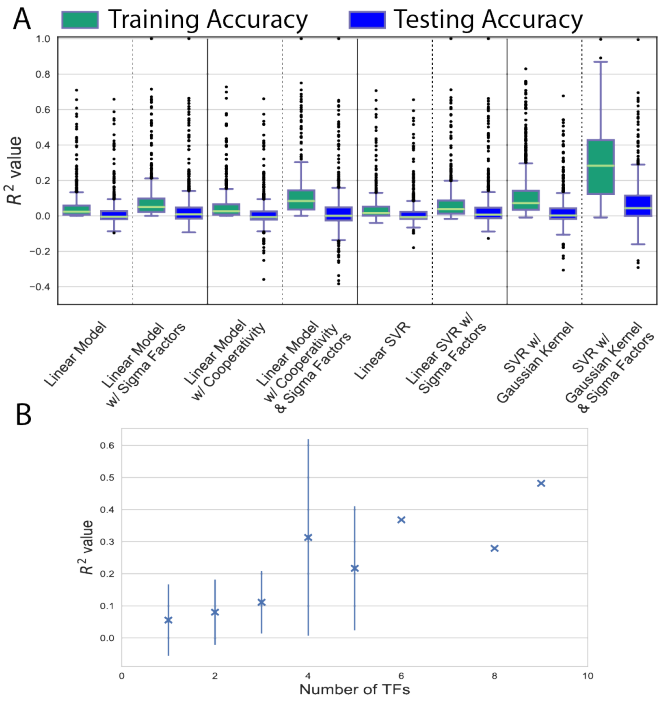

Fig. S11. Accuracy of expression predictions on training and held-out testing transcription units using only strong interactions from RegulonDB. (A) R^2 of predicted expression profile vs. true expression profile using various regression models from COLOMBOS dataset. (B) R^2 value of the testing dataset predicted by a gaussian kernel SVR, grouped by number of known TFs. Error bars indicate standard deviation for groups with $>$3 observations.

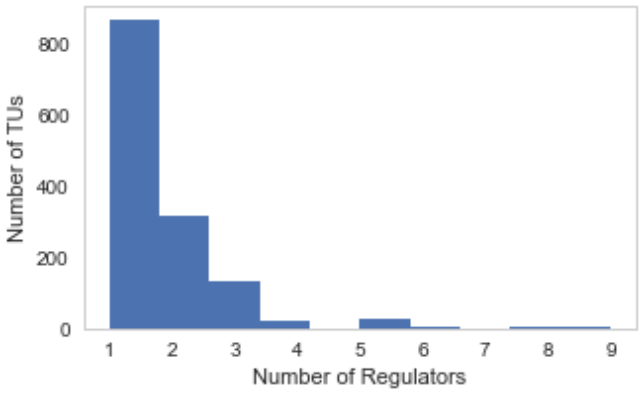

Fig. S12. Number of known regulators per TU.

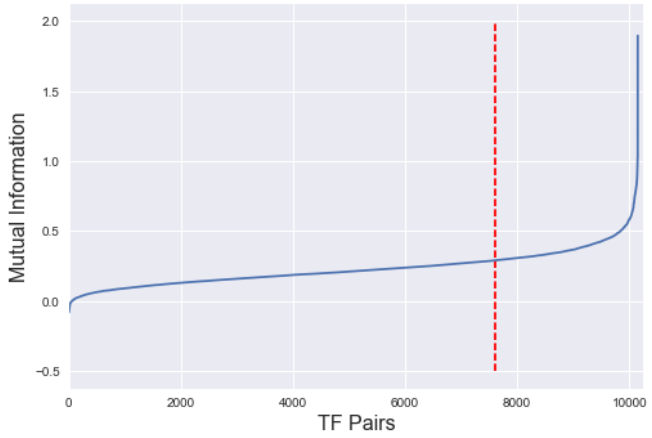

Fig. S13. Sorted MI values between all pairs of transcription factors. Any TF with mutual information higher than the 75th percentile (red dashed line) with any known regulators of each TU were prohibited from being selected as a random regulator when comparing regression results of a randomized TRN to the known TRN.

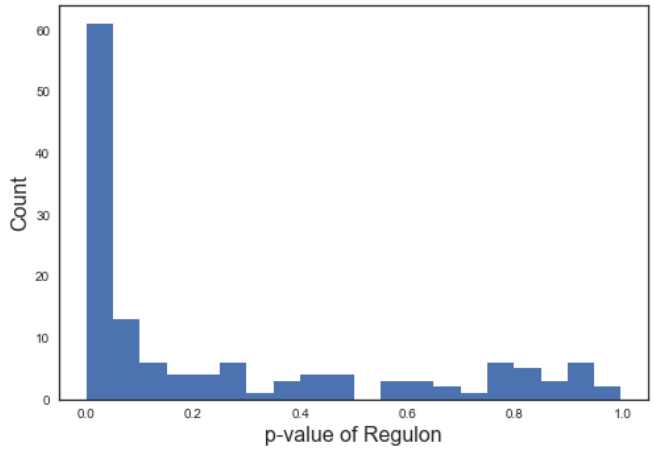

Fig. S14. P-values of observing higher mutual information between a TF and its regulon as compared to null MI distributions. The p-value was calculated by comparing the mutual information between the TF and its regulated genes with the mutual information between the TF and non-regulated genes.

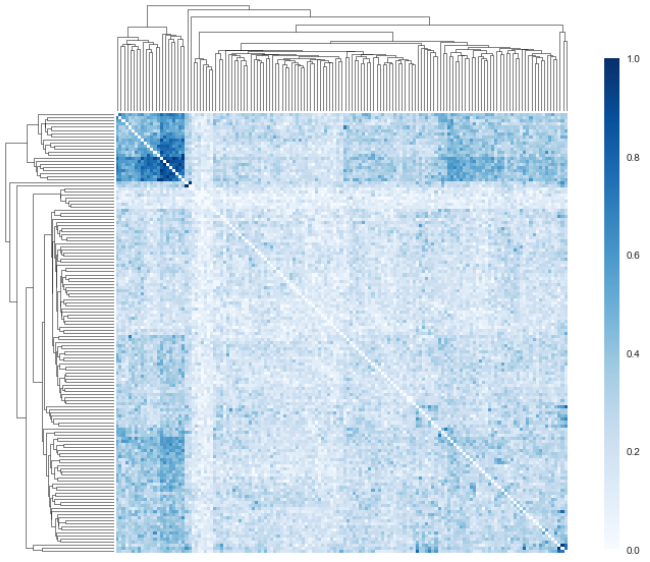

Fig. S15. Clustered heatmap of mutual information between all 147 transcription factors in the hiTRN.

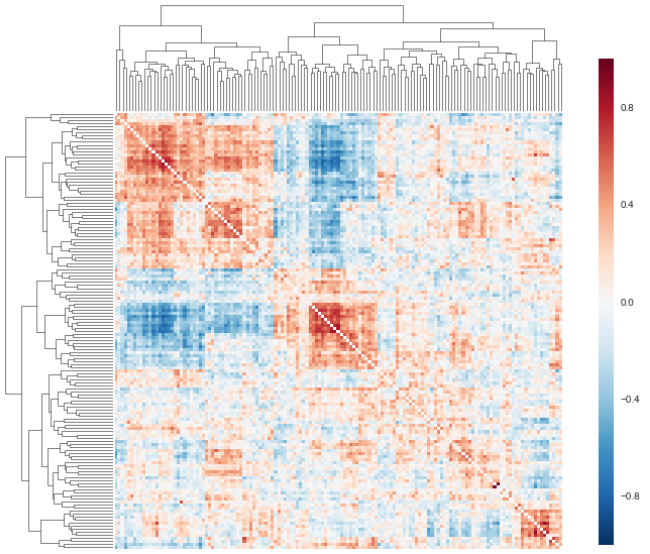

Fig. S16. Clustered heatmap of Pearson R correlation between all 147 transcription factors in the hiTRN.

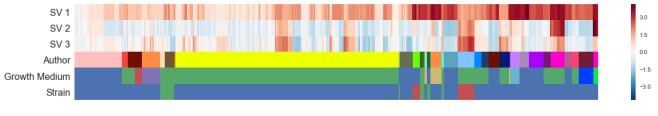

Fig. S17. Surrogate variables identified from SVA, aligned with various possible sources of expression heterogeneity, including author, growth medium and strain.

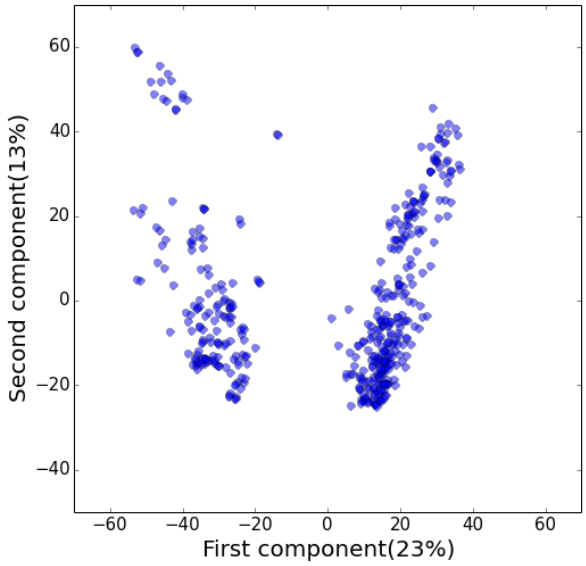

Fig. S18. The first and second principal components of the PCA analysis. Each data point represents an experiment. Data points are separated into 3 distinct groups by the first and second components.

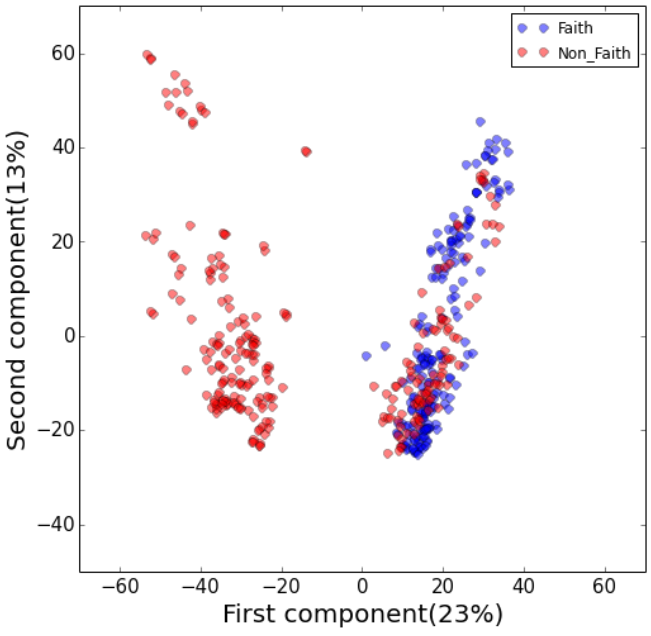

Fig. S19. The first and second principal components labeled by experimenters. The experiments done by the Faith lab and other labs are represented by blue and red dots, respectively.

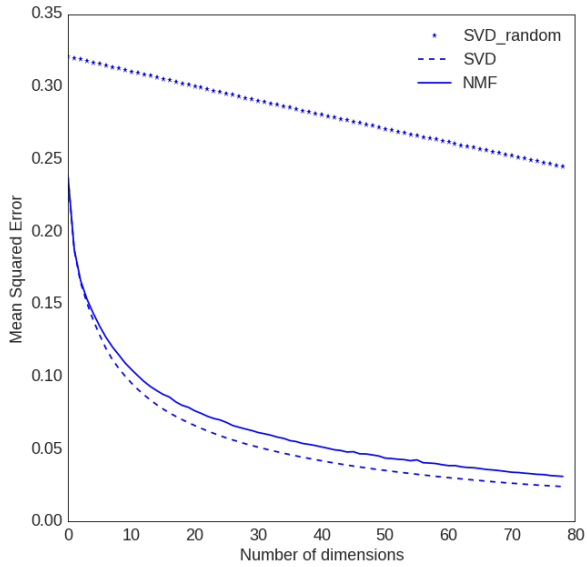

Fig. S20. Mean squared error of NMF and SVD of the EcoMAC dataset as a function of number of dimensions. SVD was also performed on random datasets for comparison. Dimensions between 35-50 are appropriate to describe the dimension of the data, as the slope of the NMF graph is similar to that of SVD on random matrix.

SI Tables

Table S1. Number of differentially expressed genes that can be traced back to the knocked out TFs

Experiment	# of DEG reachable in RegulonDB	# of DEG reachable in hiTRN
narL+Nitrate	20	20
narL/narP+Nitrate	19	19
narP+Nitrate	1	1
narL+NO	5	5
narL/narP+NO	9	9
narP+NO	1	1
arcA	**12**	**141**
fnr	**4**	**148**
arcA/fnr	0	0
oxyR+fumarate	3	6
oxyR+Nitrate	9	15
arcA	**7**	**140**
fnr	**1**	**169**
arcA/fnr	**7**	**165**
oxyR	2	3
soxS	3	3
crp+nor	30	54
dnaA+nor	0	0
fis+nor	18	27
purR	3	12
purR+adenine	1	14