

Query large scale microarray compendium datasets using a model-based Bayesian approach with variable selection

Ming Hu and Zhaohui S. Qin

Center for Statistical Genetics, Department of Biostatistics, School of Public Health, University of Michigan, Ann Arbor, MI 48109-2029.

Supplementary Material

Table of Content

- 1.** Bayesian Statistical Methods
- 2.** Detailed protocol of microarray data analysis procedure using BEST
- 3.** Query result from six other transcription factors
- 4.** **Table S1.** Performance comparison using area under the curve (AUC) of Receiver Operating Characteristic (ROC) among various methods for querying simulated microarray gene expression datasets.
- 5.** **Table S2.** Information of the 28 potential Lrp target genes identified by BEST when applied to the 100-gene test set selected from the *E. coli* microarray compendium.
- 6.** **Table S3.** Information on the 305 unique experimental conditions (612 different chips with replicates).
- 7.** **Table S4.** Information on the 57 potential Lrp target genes identified by BEST in the 300-gene test set extracted from the *E. coli* compendium.
- 8.** **Table S5.** Information on the 27 potential PdhR target genes identified by BEST in the 100-gene test set extracted from the *E. coli* compendium.
- 9.** **Table S6.** Information on the 31 potential FecI target genes identified by BEST in the 100-gene test set extracted from the *E. coli* compendium.
- 10.** **Table S7.** Information on the 31 potential LexA target genes identified by BEST in the 100-gene test set extracted from the *E. coli* compendium
- 11.** **Table S8.** Information on the 54 potential FlhC target genes identified by BEST in the 200-gene test set extracted from the *E. coli* compendium
- 12.** **Table S9.** Information on the 67 potential FlhD target genes identified by BEST in the 200-gene test set extracted from the *E. coli* compendium
- 13.** **Table S10.** Information on the 56 potential FliA target genes identified by BEST in the 200-gene test set extracted from the *E. coli* compendium
- 14.** **Figure S1.** ROC curves for various query methods when applying to synthetic datasets simulated under different settings and when there are 100% foreground columns
- 15.** **Figure S2.** ROC curves for various query methods when applying to synthetic datasets simulated under different settings and when there are 75% foreground columns.
- 16.** **Figure S3.** ROC curves for various query methods when applying to synthetic

datasets simulated under different settings and when there are 50% foreground columns.

17. **Figure S4.** Log-likelihood trace plots of the ten parallel chains resulted from the BEST run on 100-gene and 300-gene test sets selected from the *E. coli* microarray compendium.
18. **Figure S5.** Sequence logo plot (Schneider and Stephens 1990) and position specific weight matrix (PSWM) for the motif of transcription factor Lrp.
19. **Figure S6.** Boxplots of Pearson correlation coefficients.
20. **Figure S7. A.** Histograms of expression profile differences (z_{ij}) in the top five experimental conditions (foreground). **B.** Histogram of expression profile differences (z_{ij}) in the bottom five experimental conditions (background). Data used here is the 100-gene test set selected from the *E. coli* microarray compendium.
21. **Figure S8.** Trace plots of 24 predicted Lrp target genes identified by BEST that are not in the RegulonDB target set. Black lines indicate the query gene—Lrp, the red line indicate the potential target genes. Only the 139 foreground experimental conditions identified by BEST were shown in these plots.
22. **Figure S9. A.** Histograms of expression profile differences (z_{ij}) in the top five experimental conditions (foreground). **B.** Histogram of expression profile differences (z_{ij}) in the bottom five experimental conditions (background). Data used here is the 300-gene test set selected from the *E. coli* microarray compendium.
23. **Additional references.**

1. Bayesian Statistical Methods

The Bayesian Expression Search Tool (BEST) algorithm proposed in the manuscript is a Markov chain Monte Carlo (MCMC)-based computational method (Liu 2001) which assumed an explicit statistical model that describes the relationship between the query gene and the entire microarray compendium. In this section, we first describe the general Bayesian inference procedure and then the detail implementation for BEST.

The full process of a typical Bayesian analysis can be described as consisting of three main steps (Gelman et al. 1995). (a) setting up a full probability model, the joint distribution, that captures the relationship among all the variables (e.g., observed data, missing data, unknown parameters) in consideration; (b) summarizing the findings for particular quantities of interest by appropriate posterior distributions, which is typically a conditional distribution of the quantities of interest given the observed data; and (c) evaluating the appropriateness the model and suggesting improvements (model criticism and selection).

A standard procedure for carrying out step (a) is to formulate the scientific question of interest though the use of a probabilistic model, from which we can write down the likelihood function of unknown parameter. In BEST, the input data for statistical inference is the difference, $z = (z_1, z_2, \dots, z_N)^t$, between the particular query expression profile and the expression profile of genes in the database. The unknown parameters $\Theta = (R, E)$ are a row indicator vector $R = (r_1, r_2, \dots, r_N)$ and a column indicator vector $E = (e_1, e_2, \dots, e_M)$. Here R indicates whether the target genes are functionally related to the query gene, and E indicates whether co-expression occurs under the experimental conditions. We assume that the differences between a related gene and the query gene at the foreground columns follow normal distributions, and others follow a background normal distribution. So the query problem could be viewed as statistical inference from a Gaussian mixture model. Then a prior distribution $f_0(\theta)$ is contemplated, which should be both mathematically tractable and scientifically meaningful. In BEST, we adopt standard conjugate priors for these model parameters (Gelman et al. 1995). The joint probability distribution can then be represented as *Joint = likelihood × prior*, i.e.,

$$p(y, \theta) = p(y | \theta)f_0(\theta)$$

Step (b) is completed by obtaining the posterior distribution through the application of Bayes theorem:

$$p(\theta | y) = \frac{p(y, \theta)}{p(y)} = \frac{p(y | \theta)f_0(\theta)}{\int p(y | \theta)f_0(\theta)d\theta} \propto p(y | \theta)f_0(\theta)$$

After integrating out nuisance parameters, we get full conditional distributions of R and E , which are Bernoulli distributions. Finding the maximum likelihood estimator seems a good way to solve this problem, i.e., finding the genes which are most likely functionally related to the query gene, and the experimental conditions where co-expression most likely occurs. However, it is impossible to enumerate all possible combinations due to the large numbers of genes in database and so many different experimental conditions. We therefore employ a MCMC strategy to find a near-optimal solution. MCMC refers a collection of stochastic simulation techniques that can be used to sample from complicated probability distributions.

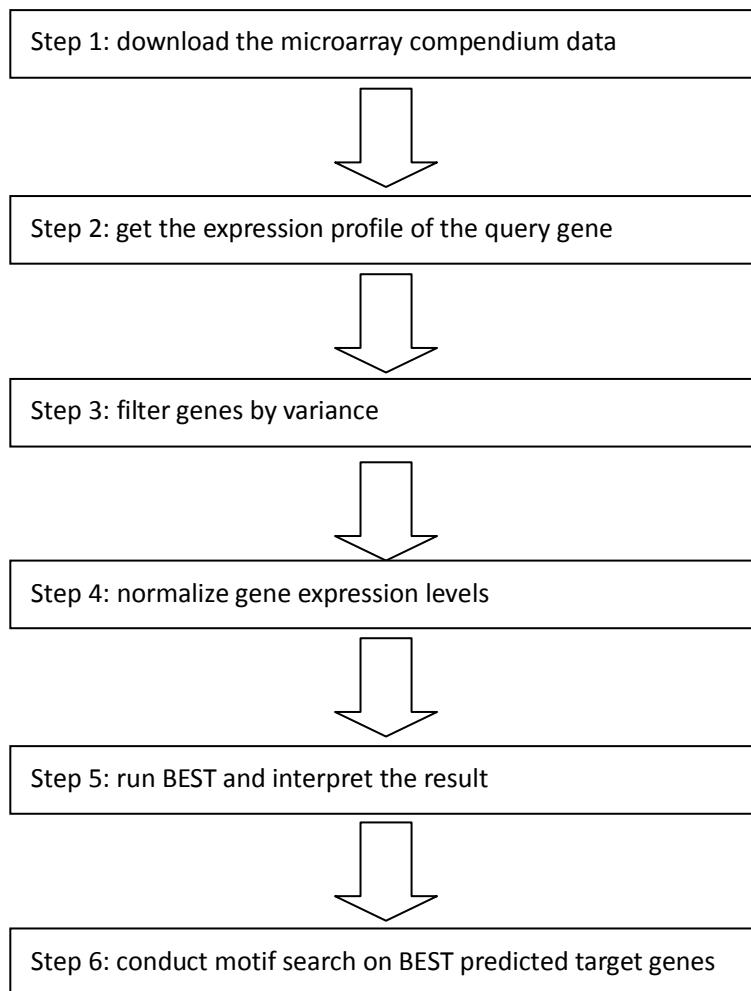
Its basic principle is to design a Markov transition rule so that the equilibrium distribution of the Markov chain is the desired distribution. One of the most popular algorithms is the Gibbs sampler, which updates one component at a time. Each component-wise update is achieved by a sample drawn from its conditional distribution with values of other components held fixed. In BEST, we use Gibbs sampler to sample joint distribution of row indicator vector $R = (r_1, r_2, \dots, r_N)$ and column indicator vector $E = (e_1, e_2, \dots, e_M)$. We draw Bernoulli sample of each indicator according to the Bayes factor between two normal distributions, with all the other indicators held constant. The posterior likelihood converges after a number of iterations, and then we report the parameters corresponding to the posterior mode as our parameter estimators. Several parallel chains are used simultaneously to prevent trapping in the local mode.

The Bayesian approach has at least two advantages. First, through the prior distribution we can use prior knowledge and information about the value of unknown parameters. This is especially important since biologists often have substantial knowledge about the subject under study. To the extent that this information is correct, it will sharpen the inference about the unknown parameters. In BEST, users can customize the parameters for prior distributions such that certain genes and experimental conditions will have a higher chance to be selected as target gene/foreground conditions. Users can even fix some indicator variables based on previous knowledge/experience. Correctly assigned informative priors can make BEST converge much faster and generate much more accurate predictions.

Second, treating all the variables in the system as random variables greatly clarifies the methods of analysis. It follows from the basic probability theory (Bayes formula) that information about the realized value of any random variable based on observation of related random variables is summarized in the conditional distribution. In BEST, we can get the close form of posterior distribution since we adopt standard conjugate priors. In addition, it is easy to integrate out nuisance parameters and results in simple distribution (Bernoulli distribution) for parameters of interest.

2. Detailed protocol of microarray data analysis procedure using BEST

The *E. coli* dataset originally came from the study reported in Faith et al. (2007). The authors conducted a comprehensive survey of gene expression profiles of all *E. coli* genes using 612 Affymetrix GeneChip arrays treated with 305 different experimental conditions. RMA normalized data (Faith et al., 2008) was used in this study, which consisted of 4,217 genes and 305 samples. The detail of microarray data analysis procedure, such as microarray profiling, bacterial strains, steady-state experiments, time-course experiments, preparation of RNA and hybridization, external data, microarray normalization, are available at (Faith et al., 2007).



Step 1: download microarray compendium data file “E_coli_v4_Build_4_norm.tar.gz” from <http://m3d.bu.edu/norm/?C=M;O=A>. This zipped data file describes the normalized compendium dumps from M3D, which contains six files with expression data. “avg_E_coli_v4_Build_4_exps305probes4217.tab”, the expression data file which contains 305 experimental conditions and 4217 genes, was used in our study.

Step 2: get the expression profile of the query gene, for example: Lrp, from the microarray compendium “avg_E_coli_v4_Build_4_exps305probes4217.tab”, which is the expression profile of Lrp across the 305 different experimental conditions.

Step 3: filter genes based on their variances. First, we calculated the variances of all 4217 genes found in the microarray compendium. We then remove all genes whose variation across all experimental conditions is less than the query gene. This purpose of filtering is to reduce computation time and to maximize the chance of finding biological meaningful targets. For the query gene Lrp, there are 524 genes (out of 4217 in total, 12%) with total expression variance greater than that of Lrp. We thus used these 524 candidate genes in our search.

Step 4: normalize the query gene and the 524 candidate genes. First, we calculated the mean and standard deviation across the 305 experimental conditions for each gene, and then normalize each of the gene expression levels by subtracting its mean and dividing by its standard deviation. After normalization, the query gene Lrp and the 524 candidate genes all have the same mean and variance (mean=0 and standard deviation=1).

Step 5: run BEST on the normalized gene expression levels using user-specified parameters such as the number of iteration in MCMC and the number of parallel chains.

Step 6: conduct motif search. We download position specific weight matrices (PSWM) from RegulonDB (http://regulondb.ccg.unam.mx/data/Matrix_AlignmentSet.txt), and the complete *E. coli* genome from GenBank

(ftp://ftp.ncbi.nih.gov/genbank/genomes/Bacteria/Escherichia_coli_K12_substr_MG1655/U_00096.fna). We then calculated the log likelihood ratio comparing between the motif model and the background model on each possible start location in the intergenic regions (up to 500 bp upstream) of genes identified by BEST. The locations with log likelihood ratio higher than a certain threshold are treated as putative motifs.

3. Query result from six other transcription factors

In addition to TF Leucine-responsive Regulatory Protein (Lrp), we used another six TFs (PdhR, FecI, LexA, FlhC, FlhD and FliA) as the query genes in this study to test BEST's performance. All six TFs have an almost equal number of target genes in ReglonDB and CLR prediction with an estimated 60% precision (Faith et al. 2007).

3.1. Query result from PdhR

PdhR has five target genes in ReglonDB. CLR predicted four target genes with 60% precision. None of them is in the RegulonDB target set. We included these nine genes and another 91 negative genes to form a 100 gene test set. BEST found 27 target genes and 179 experimental conditions as foreground. Twenty-seven of BEST's target genes included two target genes in ReglonDB and four target genes predicted by CLR (p-value of 0.0110). We found three genes (uspE, cspD, aceA) with inversed pattern. Table S4 lists all PdhR target genes identified by BEST.

3.2. Query result from FecI

FecI has six target genes in ReglonDB. CLR predicted eight target genes with 60% precision. Eight of these nine predictions are not in the RegulonDB target set. We included these 13 genes and another 87 negative genes to form a 100 gene test set. BEST found 31 target genes and 169 experimental conditions as foreground. Thirty-one of BEST's target genes included all 13 target genes in ReglonDB and target genes predicted by CLR (p-value of 2.9×10^{-8}). We found no gene with inversed pattern. Table S5 lists all FecI target genes identified by BEST.

3.3. Query result from LexA

LexA has 16 target genes. CLR predicted 17 targets genes with 60% precision. 10 of these 17 predictions are not in the RegulonDB target set. We included these 26 genes and another 74 negative genes to form a 100 gene test set. BEST found 31 target genes and 237 experimental conditions as foreground. Thirty-one of BEST's target genes included 10 target genes in ReglonDB and all target genes predicted by CLR (p-value of 1.5×10^{-8}). We found one gene (uspE) with inversed pattern. Table S6 lists all LexA target genes identified by BEST.

3.4. Query result from FlhC

FlhC has 30 target genes in ReglonDB. CLR predicted 53 targets genes with 60% precision. 24 of these 53 predictions are not in the RegulonDB target set. We included these 54 genes and another 146 negative genes to form a 200 gene test set. BEST found 54 target genes and 266 experimental conditions as foreground. Fifty-four of BEST's target genes included 29 target genes in ReglonDB and all target predicted by CLR (p-value of 2.7×10^{-46}). We found no gene with inversed pattern. yjdA is the new hypothetical FlhC target gene identified by BEST in addition to false positive genes in Faith's prediction with 60% precision. Table S7 lists all FlhC target genes identified by BEST.

3.5. Query result from FlhD

FlhD has 46 target genes in ReglonDB. CLR predicted 46 target genes with 60% precision. Twenty of these 46 predictions are not in the RegulonDB target set. We included these 66

genes and another 134 negative genes to form a 200 gene test set. BEST found 55 target genes and 215 experimental conditions as foreground. Fifty-five of BEST's target genes included 29 target genes in ReglonDB and all target genes predicted by CLR (p-value of 1.67×10^{-17}). We found two genes (*micF*, *gadX*) with inverted pattern. *cheY*, *cheZ*, *flxA*, *micF*, *gadX* and *yjdA* are the six new hypothetical FlhD target genes identified by BEST in addition to false positive genes predicted by CLR with 60% precision. Table S8 lists all FlhC target genes identified by BEST.

3.6. Query result from FliA

FliA has 42 target genes in ReglonDB. CLR predicted 56 target genes with 60% precision. Fifteen of these 56 predictions are not in the RegulonDB target set. We included these 57 genes and another 143 negative genes to form a 200 gene test set. BEST found 56 target genes and 281 experimental conditions as foreground. Fifty-six of BEST's target genes included 41 genes in ReglonDB and all target predicted by CLR (p-value of 4.08×10^{-47}). We found no genes with inverted pattern, and no new hypothetical FliA target gene identified by BEST in addition to false positive genes predicted by CLR with 60% precision. Table S9 list all FliA target genes identified by BEST.

Table S1. Performance comparison using area under the curve (AUC) of Receiver Operating Characteristic (ROC) among various methods for querying simulated microarray gene expression datasets. Best results are displayed in bold.

Case	Sub-case [*]	Pearson ^a	Spearman ^b	Kendall ^c	QDB ^d	Mutual ^e	BEST A ^f	BEST B ^g	BEST C ^h
Case 1: 100% foreground	I	1	1	1	1	1	1	1	1
	II	0.68	0.68	0.68	0.85	1	1	1	1
	III	1	1	1	1	1	1	1	1
	IV	0.66	0.71	0.71	0.79	0.85	1	1	1
Case 2: 75% foreground	I	0.98	1	1	1	0.98	1	1	1
	II	0.7	0.71	0.71	0.89	0.95	1	1	1
	III	0.99	1	1	1	0.98	1	1	1
	IV	0.67	0.71	0.71	0.84	0.88	1	1	1
Case 3: 50% foreground	I	0.89	0.93	0.95	0.99	0.87	1	1	1
	II	0.67	0.69	0.69	0.89	0.81	1	1	1
	III	0.88	0.92	0.94	0.92	0.84	1	1	1
	IV	0.59	0.62	0.64	0.77	0.69	0.98	0.99	0.99
Case 4: 25% foreground	I	0.68	0.69	0.71	0.55	0.61	0.91	0.98	0.99
	II	0.55	0.55	0.56	0.51	0.58	0.85	0.92	0.99
	III	0.67	0.68	0.69	0.51	0.59	0.93	0.98	0.99
	IV	0.54	0.55	0.56	0.53	0.51	0.79	0.84	0.88

* There are four sub-cases in each of the simulated cases with the same amount of foreground columns.

Sub case I: no linear transformation;

Sub case II: only add linear transformation;

Sub case III: only add cell-level noise;

Sub case IV: add both linear transformation and cell-level noise.

^aQuery method using Pearson correlation coefficient.

^bQuery method using Spearman correlation coefficient.

^cQuery method using Kendall's τ .

^dQuery method using QDB.

^eQuery method using mutual information.

^fQuery method using BEST.

^gQuery method using BEST allowing exclusion of individual cells from the foreground.

^hQuery method using BEST when fixing the indicator variables of five true target genes and five true experimental conditions as 1.

Table S2. Information of the 28 potential Lrp target genes identified by BEST when applied to the 100-gene test set selected from the *E. coli* microarray compendium.

Rank	Gene Name ^a	Log Bayes Ratio	positive/negative ^b	RegulonDB ^c	CLR ^d
1	serA	131.81		X	X
2	leuA	129.99		X	X
3	leuL	128.72		X	X
4	gltD	128.22		X	X
5	leuD	123.19		X	X
6	ilvI	120.44		X	
7	ilvH	119.61		X	X
8	gltB	119.04		X	
9	leuC	116.22		X	X
10	livG	115.37		X	X
11	ilvE	114.30		X	
12	livK	113.30		X	X
13	leuB	110.28		X	X
14	livJ	109.72		X	
15	livM	108.48		X	
16	gcvB	107.80	negative		
17	serC	103.20		X	X
18	aroA	97.58		X	X
19	livH	94.76		X	X
20	livF	93.82		X	X
21	ilvL	90.43		X	
22	ilvD	89.88		X	
23	lysU	84.52	negative	X	
24	tbl	81.47	negative	X	
25	tdh	80.09	negative	X	
26	ilvG	79.47		X	
27	ilvM	71.23		X	X
28	ilvA	65.02		X	

^a Genes displayed here are sorted by the Log Bayes ratio (target gene versus non-target gene).

^b Blank indicates that the target gene shows the same pattern as the query gene.

Negative indicates that the target gene shows the inversed pattern as the query gene.

^c BEST identifies 27 genes among 61 target genes in RegulonDB. “X” indicates that the predicted gene is in the RegulonDB target set.

^d “X” indicates that the gene is predicted by CLR as a target gene.

Table S3. Information on the 305 distinct experimental conditions (among 612 different chips with replicates). The 305 experimental conditions are sorted by log Bayes ratio. BEST predicts the top 143 as foreground and the rest as background. Detail information of these 612 chips and 305 experimental conditions could be found in microarray compendium data file “E_coli_v4_Builid_4_norm.tar.gz” from the Gardner Lab. (<http://m3d.bu.edu/norm/?C=M;O=A>)

Rank	Experimental Conditions	Replicates	Log Bayes Ratio
1	M9_K_arcA_anaerobic	3	50.47
2	M9_WT	3	48.79
3	M9_K_appY_anaerobic	3	44.81
4	M9_K_soxS_anaerobic	3	43.21
5	M9_K_arcA	3	38.07
6	M9_K_oxyR_anaerobic	3	37.60
7	M9_K_soxS	3	37.23
8	M9_K_arcAfnr	3	37.02
9	M9_WT_anaerobic	4	35.75
10	M9_K_oxyR	3	35.52
11	lacZ_W1863_t0	1	35.24
12	M9_K_fnr	3	34.34
13	ccdB_MG1655_t30	2	34.14
14	M9_K_fnr_anaerobic	3	34.05
15	ccdB_W1872_t60	1	31.42
16	lacZ_W1863_t60	1	31.11
17	ccdB_W1872_t30	1	30.66
18	ccdB_W1872_t90	1	29.37
19	M9_K_appY	3	29.01
20	lacZ_MG1063_t0	2	27.61
21	ccdB_MG1655_t0	2	26.92
22	norfloxacin_chelator_MG1063_t0	1	26.75
23	ccdB_chelator_MG1063_t0	1	26.31
24	ccdB_chelator_MG1063_t60	1	26.12
25	lacZ_W1863_t30	1	25.47
26	fnr_K_fnrAnaerobic	4	25.21
27	cybr_N_log	2	24.95
28	MG1063_uninduced_t0	1	24.50
29	MG1063_uninduced_t60	1	24.36
30	MG1063_uninduced_t30	1	24.22
31	suspension_7hr	1	23.49
32	MG1655_ampicillin_t120	1	23.40
33	ccdB_chelator_W1872_t0	1	23.30
34	MG1655_ampicillin_t30	1	23.23
35	ik_H2_T3	1	23.04

36	fnr_wtAnaerobic	3	22.85
37	ik_L2_T3.5	1	22.80
38	norfloxacin_BW25113_t120	1	22.59
39	norfloxacin_BW25113recA_t60	1	22.39
40	har_S1_R_noIPTG	3	21.62
41	ccdB_chelator_MG1063_t30	1	21.60
42	cybr_O_log	2	21.56
43	biofilm_4hr	1	21.32
44	ph5.7_anaerobic	5	21.24
45	carbonSourceForaging	2	21.02
46	WT_MOPS_glucose	5	20.77
47	ik_L2_T4	1	20.38
48	M9_K_arcAfnr_anaerobic	3	20.28
49	WT_MOPS_glycerol	2	20.08
50	ik_H2_T3.5	1	19.86
51	MG1655_uninduced_t0	1	19.36
52	luc2_U_N0000	2	19.30
53	MG1655_ampicillin_t60	1	18.85
54	ccdB_BW25113recA_t120	1	18.84
55	ik_L2_T3	1	18.77
56	menC_U_N0075	3	18.76
57	crcB_U_N0075	3	18.69
58	MG1655_t480_aerobic	2	18.61
59	cpxR_U_N0075	3	18.31
60	era_U_N0075	3	18.26
61	crp_U_N0075	3	18.09
62	luc_U_N0075	3	18.00
63	gcvR_U_N0075	3	17.83
64	dnaA_U_N0075	3	17.79
65	menB_U_N0075	3	17.65
66	fis_U_N0075	3	17.62
67	ccdB_BW25113recA_t30	1	17.60
68	ccdB_MG1063_t0	2	17.18
69	ik_H2_T2.5	1	17.04
70	WT_N0075	2	17.04
71	ccdB_W1872_t0	1	16.89
72	MG1655_t720_aerobic	2	16.86
73	norfloxacin_BW25113recA_t30	1	16.59
74	MG1655_t270_anaerobic	2	16.35
75	lacZ_W1863_t90	1	16.27
76	rimI_U_N0075	3	16.19
77	minD_U_N0075	3	16.15
78	zipA_U_N0075	3	15.99
79	dinP_U_N0025	3	15.75

80	lacZ_MG1063_t30	2	15.74
81	mcrB_U_N0075	3	15.62
82	ccdB_BW25113recA_t180	1	15.53
83	MG1655_uninduced_t60	1	15.48
84	norfloxacin_chelator_MG1063_t0.1	1	15.34
85	yebF_U_N0075	3	15.21
86	MG1655_t150_aerobic	2	15.13
87	ccdB_BW25113_t180	1	15.01
88	biofilm_wt_noGlucose	1	14.83
89	MG1655_t225_anaerobic	2	14.75
90	ccdB_BW25113_t30	1	14.61
91	ast_pBADsup2	3	14.50
92	ph7	5	14.17
93	recA_U_N0025	3	13.88
94	MG1655_t300_aerobic	2	13.71
95	uspA_U_N0075	3	13.53
96	norfloxacin_MG1063_t0	1	13.47
97	mcrC_U_N0075	3	13.41
98	bcp_U_N0075	3	13.22
99	b2618_U_N0075	3	13.19
100	pepAA_t0	2	13.00
101	nupC_U_N0075	3	12.84
102	ldrA_U_N0075	3	12.34
103	ccdB_K12_t90	1	12.15
104	lacZ_K12_t60	1	12.09
105	BW25113_uninduced_t120	1	11.80
106	yoeB_U_N0075	3	11.71
107	minE_U_N0075	3	11.48
108	ph7_anaerobic	5	11.45
109	luc_U_N0025	3	11.44
110	ast_pBAD18	3	11.42
111	MG1655_t405_aerobic	2	11.28
112	sbcB_U_N0075	3	10.53
113	fklB_U_N0075	3	10.45
114	lacZ_MG1655_t0	1	10.36
115	MG1655_t1080_aerobic	2	9.96
116	gyrA_U_N0075	3	9.66
117	har_S0_R_noIPTG	5	9.52
118	norfloxacin_BW25113_t180	1	9.51
119	suspension_4hr	1	9.45
120	W3110_K_luxS	2	9.23
121	murI_U_N0075	3	8.80
122	MG1655_uninduced_t120	1	8.43
123	BW25113recA_uninduced_t180	1	7.97

124	folA_U_N0075	3	7.76
125	lacZ_K12_t30	1	7.69
126	cspF_U_N0075	3	7.61
127	dnaN_U_N0075	3	6.63
128	ccdB_BW25113recA_t0	1	6.57
129	W3110_wt	2	6.43
130	ccdB_K12_t30	1	5.91
131	hlpA_U_N0075	3	5.48
132	MG1655_uninduced_t30	1	4.20
133	biofilm_7hr	1	3.80
134	T60_N10000	3	3.16
135	MG1655_t180_anaerobic	2	2.78
136	WT_N0000	2	2.69
137	dinI_U_N0025	3	0.68
138	BW25113recA_uninduced_t120	1	0.29
139	ccdB_BW25113recA_t60	1	0.29
140	ik_L2_T4.5	1	0.22
141	dam_U_N0075	3	0.05
142	lacZ_MG1655_t60	1	-0.22
143	rstB_U_N0075	3	-0.71
144	pepAA_t30	2	-1.03
145	dnaT_U_N0075	3	-1.16
146	MGD1_t30	2	-1.26
147	norfloxacin_BW25113recA_t0	1	-1.62
148	gyrI_U_N0075	3	-1.83
149	MG1655_kanamycin_t120	1	-1.95
150	ik_H2_T8	1	-2.05
151	sulA_U_N0025	3	-2.29
152	umuD_U_N0025	3	-2.38
153	ccdB_K12_t60	1	-2.44
154	T0_N0000	3	-2.50
155	ik_H2_T4.5	1	-2.56
156	norfloxacin_BW25113_t30	1	-2.93
157	lacZ_MG1655_t90	1	-2.93
158	ik_H2_T6	1	-3.02
159	ik_L2_T2.5	1	-3.04
160	ruvA_U_N0025	3	-3.06
161	ccdB_BW25113_t0	1	-3.09
162	lexA_U_N0025	3	-3.54
163	ik_L2_T5.5	1	-3.60
164	holD_U_N0075	3	-3.87
165	nrdA_U_N0075	3	-4.06
166	ccdB_K12_t120	1	-4.15
167	WT_D_N0100	2	-4.29

168	pepCO_t30	2	-4.37
169	lon_U_N0025	3	-4.56
170	recA_D_N0100	2	-4.60
171	MG1655_kanamycin_t60	1	-4.62
172	MG1655_kanamycin_t30	1	-4.65
173	ccdB_BW25113_t60	1	-4.90
174	galF_U_N0075	3	-5.04
175	ruvC_U_N0075	3	-5.24
176	T48_N10000	3	-5.43
177	MG1063_uninduced_t120	1	-5.44
178	ccdB_chelator_W1872_t30	1	-5.66
179	lacZ_MG1063_t90	2	-5.84
180	BW25113recA_uninduced_t0	1	-5.95
181	uvrA_U_N0025	3	-6.13
182	IHF_U_N0075	2	-6.15
183	MOPS_K_dps_stationary2	1	-6.15
184	relA_U_N0025	3	-6.27
185	BW25113recA_uninduced_t30	1	-7.01
186	emrR_U_N0075	2	-7.21
187	ik_H2_T4	1	-7.22
188	cybr_O	2	-7.24
189	WT_MOPS_stationary3	2	-7.39
190	WT_MOPS_heatShock	1	-7.60
191	MOPS_K_dps_stationary	2	-7.61
192	cybr_N	2	-7.66
193	luc_U_N0000	3	-7.72
194	ph8.5_anaerobic	5	-7.86
195	pyrC_U_N0075	3	-8.00
196	ph5	5	-8.10
197	BW25113_uninduced_t30	1	-8.33
198	cybr_N_stat	2	-8.51
199	WT_MOPS_stationary2	2	-8.53
200	hscA_U_N0075	3	-8.60
201	ik_L2_T6	1	-8.67
202	ik_H2_T5.5	1	-8.94
203	mazF_U_N0025	3	-9.11
204	pET3d_t0	2	-9.19
205	luc2_U_N0025	2	-9.21
206	T24_N10000	3	-9.22
207	nrdB_U_N0075	2	-9.36
208	BW25113_uninduced_t180	1	-9.42
209	T36_N10000	3	-9.74
210	MG1655_spectinomycin_t30	1	-9.74
211	norfloxacin_BW25113_t60	1	-9.75

212	WT_MOPS_cipro2	1	-10.61
213	ccdB_MG1655_t60	2	-10.66
214	MGD1_t0	2	-10.98
215	WT_MOPS_acetate	2	-10.99
216	ccdB_chelator_MG1063_t120	1	-11.15
217	ccdB_MG1063_t30	2	-11.39
218	lacZ_K12_t120	1	-11.41
219	cybr_O_stat	2	-11.44
220	W3110_K_luxS_glucose	1	-11.44
221	lacZ_MG1063_t60	2	-11.71
222	ccdB_BW25113_t120	1	-11.87
223	ccdB_MG1655_t90	2	-11.91
224	ik_L2_T5	1	-12.14
225	WT_MOPS_proline	2	-12.22
226	recA_D_N0050	2	-12.34
227	WT_MOPS_stationary4	2	-12.42
228	WT_MOPS_cipro	1	-12.81
229	WT_N0025	2	-12.92
230	T12_N10000	3	-13.07
231	lacZ_MG1063_120	1	-13.07
232	BW25113recA_uninduced_t60	1	-13.11
233	fnr_K_fnrAerobic	3	-13.21
234	MG1655_t150_anaerobic	2	-13.42
235	norfloxacin_BW25113_t0	1	-13.55
236	recA_D_N0000	2	-13.59
237	WT_MOPS_acidShock	2	-13.77
238	MG1655_spectinomycin_t60	1	-13.82
239	MG1655_spectinomycin_t120	1	-14.22
240	biofilm_15hr	1	-14.39
241	K12_t360	3	-14.44
242	ccdB_K12_t0	1	-14.44
243	WT_MOPS_lateLog	3	-14.46
244	suspension_15hr	1	-15.01
245	W3110_wt_glucose	2	-15.23
246	har_S4_noIPTG	3	-15.91
247	norfloxacin_BW25113recA_t120	1	-16.73
248	ik_H2_T5	1	-16.96
249	har_S1_noIPTG	3	-16.97
250	norfloxacin_MG1063_t30	1	-17.00
251	lacZ_MG1655_t30	1	-17.38
252	pepCO_t0	2	-17.62
253	lacZ_K12_t0	1	-17.85
254	lacZ_K12_t90	1	-18.07
255	har_S1_IPTG	3	-18.16

256	MOPS_K_cspA	1	-18.20
257	MG1655_t86400_cecum	5	-18.47
258	BW25113_uninduced_t60	1	-18.67
259	norfloxacin_chelator_MG1063_t0.2	1	-18.95
260	norfloxacin_chelator_MG1063_t0.3	1	-19.00
261	MG1655_t1560_aerobic	2	-19.27
262	biofilm_24hr	1	-19.31
263	biofilm_K_yceP	1	-19.35
264	MG1655_norfloxacin_t30	1	-19.60
265	MG1655_norfloxacin_t120	1	-19.95
266	har_S0_noIPTG	3	-20.33
267	MOPS_K_dps	3	-20.40
268	cybr_KNO_N	2	-20.71
269	har_S4_IPTG	3	-21.49
270	ph8.7	5	-21.97
271	K12_t150_K_fis	3	-22.24
272	ik_L2_T8	1	-22.32
273	T24_N0000	3	-22.43
274	pET3d_t30	2	-22.91
275	K12_t90_K_fis	3	-23.30
276	BW25113_uninduced_t0	1	-23.45
277	K12_t360_K_fis	3	-23.71
278	biofilm_K_yceP_indole	2	-25.13
279	WT_N0050	2	-25.65
280	T60_N0000	3	-25.75
281	MOPS_K_hupB	1	-26.19
282	norfloxacin_MG1063_t60	1	-26.30
283	WT_MOPS_stationary	2	-26.59
284	K12_t240_K_fis	3	-26.70
285	K12_t150	3	-27.19
286	har_S4_R_IPTG	3	-27.85
287	norfloxacin_BW25113recA_t180	1	-29.03
288	ccdB_MG1063_t60	2	-29.70
289	MG1655_norfloxacin_t60	1	-30.08
290	ccdB_MG1063_t120	1	-31.01
291	ccdB_chelator_W1872_t60	1	-31.04
292	K12_t240	3	-32.33
293	MOPS_K_crp	3	-32.52
294	MG1063_uninduced_t180	1	-32.69
295	biofilm_K_tnaA	1	-32.85
296	MOPS_K_hns	3	-33.23
297	ccdB_chelator_W1872_t120	1	-33.23
298	suspension_24hr	1	-33.70
299	biofilm_wt_glucose	1	-34.55

300	har_S1_R_IPTG	3	-35.58
301	biofilm_K_trpE	1	-37.15
302	ccdB_MG1063_t90	2	-37.89
303	norfloxacin_MG1063_t120	1	-45.15
304	K12_t90	3	-45.89
305	har_S4_R_noIPTG	3	-51.28

Table S4. Information on the 57 potential Lrp target genes identified by BEST in the 300-gene test set extracted from the *E. coli* compendium.

Rank	Gene Name ^a	Log Bayes Ratio	positive/negative ^b	RegulonDB ^c	CLR ^d
1	serA	187.8		X	X
2	gltD	182.97		X	X
3	metE	180.95			X
4	leuL	178.03		X	X
5	leuD	175.01		X	X
6	leuA	174.43		X	X
7	gltB	173.3		X	
8	livG	172.8		X	X
9	livJ	172.44		X	
10	ilvE	172.35		X	
11	ompT [†]	169.81			X
12	pyrI	169.77			
13	livK	169.16		X	X
14	ilvH	168.41		X	X
15	leuC	168.1		X	X
16	ilvI	168		X	
17	gcvB	166.47	negative		
18	serC	164.09		X	X
19	livM	163.56		X	
20	leuB	163.18		X	X
21	pyrB	161.91			
22	yagU [†]	158.71			X
23	aroA	158.46		X	X
24	cysD	158.39			
25	ilvD	157.84		X	
26	lysU	157.56	negative	X	
27	livH	157.28		X	X
28	livF	155.74		X	X
29	stpA	153.87		X	
30	cysK	152.23			
31	pheL	151.32			
32	tnaC	149.8	negative		
33	dppA	148.7			
34	cysN	147.61			
35	kbl	145.76	negative	X	
36	treC	143.28	negative		
37	ilvL	142.44		X	
38	tdh	140.63	negative	X	
39	pyrL	140.4			
40	ilvC	139.63			

41	sdaA	138.77	negative	X
42	sdaC	136.96	negative	
43	ilvA	136.74		X
44	thrL	135.9		
45	hisL	135.55		
46	yeeD	133.82		
47	ilvM	131.85		X X
48	treB	130.2	negative	
49	ompF	129.21		X
50	fdoG	127.87	negative	
51	oppA	127.26		X
52	oppB	124.59		X
53	rmf	122.65		
54	oppF	122.24		X
55	ynaJ	118.39		
56	ilvG	113.06		X
57	sroF	109.54		

^a Genes displayed here are sorted by the Log Bayes ratio (target gene versus non-target gene).

^b Blank mean the target gene shows the same pattern as the query gene. Negative means the target gene shows the inversed pattern as the query gene.

^c BEST indentifies 33 genes among 61 target genes in RegulonDB. “X” indicates that the predicted gene is in the RegulonDB target set.

^d “X” indicates that the gene is predicted by CLR as a target gene.

[†] Previously unknown targets of Lrp, experimentally verified by ChIP (Faith et al. 2007).

Table S5. Information on the 27 potential PdhR target genes identified by BEST in the 100-gene test set extracted from the *E. coli* compendium.

Rank	Gene Name ^a	Log Bayes Ratio	positive/negative ^b	RegulonDB ^c	CLR ^d
1	recN	348.48			
2	intE	342.08			
3	recA	337.20			
4	tisB	336.43			
5	xisE	332.99			
6	araB	330.06			
7	araA	328.56			
8	sulA	328.12			
9	araD	327.70			
10	ymfJ	322.03			
11	ymfT	317.29			
12	ymfL	316.94			
13	araE	314.20			
14	murC	307.50			X
15	ftsW	306.20			X
16	murD	304.53			X
17	ndh	298.38			X
18	aceE	283.28			X
19	aceF	279.67			X
20	uspE	275.72	negative		
21	proV	274.19			
22	cspD	267.67	negative		
23	isrB	263.68			
24	spf	248.58			
25	cspA	239.93			
26	tisA	221.34			
27	aceA	76.68	negative		

^a Genes displayed here are sorted by the Log Bayes ratio (target gene versus non-target gene).

^b Blank indicates that the target gene shows the same pattern as the query gene.

Negative indicates that the target gene shows the inversed pattern as the query gene.

^c BEST identifies two genes among five target genes in RegulonDB. “X” indicates that the predicted gene is in the RegulonDB target set.

^d “X” indicates that the gene is predicted by CLR as a target gene.

Table S6. Information on the 31 potential FecI target genes identified by BEST in the 100-gene test set extracted from the *E. coli* compendium.

Rank	Gene Name ^a	Log Bayes Ratio	positive/negative ^b	RegulonDB ^c	CLR ^d
1	ymfT	206.62			
2	ymfJ	200.23			
3	ymfL	196.67			
4	araD	189.08			
5	xisE	185.50			
6	araB	185.40			
7	araA	182.57			
8	recN	177.90			
9	araE	177.45			
10	tisB	146.18			
11	tisA	144.54			
12	sulA	133.96			
13	recA	118.89			
14	proV	81.52			
15	fecE	72.40		X	
16	fecB	67.88		X	
17	isrB	63.99			
18	fecD	63.63		X	
19	fecC	63.13		X	
20	fecA	62.00		X	
21	fhuF	54.12			X
22	ybaN	53.60			X
23	exbB	46.24			X
24	fhuA	44.64			X
25	exbD	43.74			X
26	fecR	43.29		X	X
27	bfd	33.30			
28	micF	32.89			
29	spf	29.53			
30	cspA	19.14			
31	entB	-0.34			X

^aGenes displayed here are sorted by the Log Bayes ratio (target gene versus non-target gene).

^bBlank indicates that the target gene shows the same pattern as the query gene.

Negative indicates that the target gene shows the inversed pattern as the query gene.

^cBEST identifies all six genes among six target genes in RegulonDB. “X” indicates that the predicted gene is in the RegulonDB target set.

^d“X” indicates that the gene is predicted by CLR as a target gene.

Table S7. Information on the 31 potential LexA target genes identified by BEST in the 100-gene test set extracted from the *E. coli* compendium

Rank	Gene Name ^a	Log Bayes Ratio	positive/negative ^b	RegulonDB ^c	CLR ^d
1	dinF	370.20		X	X
2	araB	365.72			
3	araA	365.42			
4	araE	359.80			
5	araD	358.89			
6	ymfJ	349.63			
7	ymfL	347.87			
8	recN	347.63		X	X
9	xisE	339.91			
10	yebG	333.06			X
11	ymfT	332.53			
12	dinI	320.63			X
13	recX	318.60			
14	umuD	316.82		X	X
15	tisB	314.57			
16	yafN	314.26			X
17	tisA	312.63			
18	dinD	312.63			X
19	uvrA	310.18		X	X
20	dinG	308.19			X
21	yafO	306.66			X
22	sulA	305.85		X	X
23	polB	299.75		X	
24	recA	292.55		X	X
25	umuC	289.20		X	X
26	dinB	282.91			
27	bssS	262.57			
28	ssb	262.39		X	
29	uvrD	259.91		X	
30	yebF	242.58			X
31	uspE	239.61	negative		

^a Genes displayed here are sorted by the Log Bayes ratio (target gene versus non-target gene).

^b Blank indicates that the target gene shows the same pattern as the query gene.

Negative indicates that the target gene shows the inversed pattern as the query gene.

^c BEST identifies ten genes among 16 target genes in RegulonDB. “X” indicates that the predicted gene is in the RegulonDB target set.

^d “X” indicates that the gene is predicted by CLR as a target gene.

Table S8. Information on the 54 potential FlhC target genes identified by BEST in the 200-gene test set extracted from the *E. coli* compendium

Rank	Gene Name ^a	Log Bayes Ratio	positive/negative ^b	RegulonDB ^c	CLR ^d
1	flgE	422.41		X	X
2	fliA	417.80		X	X
3	flgC	415.00		X	X
4	flgB	412.97		X	X
5	flgG	406.95		X	X
6	flgH	406.29		X	X
7	flgD	403.29		X	X
8	flhD	401.18			X
9	motB	398.98			X
10	fliL	398.54		X	X
11	fliN	398.03		X	X
12	flgI	397.19		X	X
13	flgK	396.32			X
14	flgA	396.03		X	X
15	fliK	389.85		X	X
16	fliM	388.79		X	X
17	flgF	388.45		X	X
18	motA	387.71			X
19	cheA	386.19			X
20	cheW	385.20			X
21	fliZ	384.50		X	X
22	fliJ	384.25		X	X
23	flgM	382.36			X
24	fliF	382.00		X	X
25	flgN	380.73			X
26	fliS	380.29			X
27	cheY	378.15			X
28	flgJ	372.04		X	X
29	cheZ	371.92			X
30	cheR	371.27			X
31	yecR	370.70			X
32	cheB	369.11			X
33	fliG	367.55		X	X
34	fliC	366.59			X
35	flgL	366.38			X
36	fliH	362.55		X	X
37	fliD	360.82			X
38	fliP	356.28		X	X
39	fliQ	350.16		X	X
40	tar	347.51			X

41	fliI	345.24		X	X
42	ycgR	344.20			X
43	tap	338.41			X
44	fliE	326.37		X	X
45	flxA	325.67			X
46	fliO	324.78		X	X
47	ymdA	314.15			X
48	flhA	301.22		X	X
49	flhE	300.16		X	X
50	flhB	290.68		X	X
51	fliR	275.89		X	X
52	yhjH	272.24			X
53	tsr	255.05			X
54	yjdA	206.14			

^a Genes displayed here are sorted by the Log Bayes ratio (target gene versus non-target gene).

^b Blank indicates that the target gene shows the same pattern as the query gene.

Negative indicates that the target gene shows the inversed pattern as the query gene.

^c BEST identifies 29 genes among 30 target genes in RegulonDB. “X” indicates that the predicted gene is in the RegulonDB target set.

^d “X” indicates that the gene is predicted by CLR as a target gene.

Table S9. Information on the 67 potential FlhD target genes identified by BEST in the 200-gene test set extracted from the *E. coli* compendium

Rank	Gene Name ^a	Log Bayes Ratio	positive/negative ^b	RegulonDB ^c	CLR ^d
1	flhC	333.41			X
2	flgE	312.18		X	X
3	flgB	311.50		X	X
4	flgH	311.46		X	X
5	fliA	311.18		X	X
6	flgC	310.34		X	X
7	flgG	307.33		X	X
8	flgD	299.71		X	X
9	flgK	299.55			X
10	flgI	299.12		X	X
11	fliZ	298.07		X	X
12	flgA	296.58		X	X
13	motB	294.72			X
14	cheW	293.40			X
15	flgF	293.22		X	X
16	fliN	292.68		X	X
17	fliL	292.29		X	X
18	flgN	291.93			X
19	fliF	290.99		X	X
20	flgM	290.12			X
21	fliM	288.74		X	X
22	cheA	288.54			X
23	fliK	285.84		X	X
24	fliS	282.94			X
25	motA	282.93			X
26	yecR	282.39			X
27	flgL	281.34			X
28	fliJ	279.79		X	X
29	fliC	275.71			X
30	flgJ	275.32		X	X
31	fliD	271.73			X
32	cheR	271.62			X
33	fliP	270.94		X	X
34	fliG	270.80		X	X
35	cheB	269.55			X
36	cheY	269.30			
37	fliH	264.98		X	
38	tar	264.64			X
39	cheZ	264.38			
40	fliE	260.07		X	X

41	fliI	256.40	X	X
42	ycgR	251.77		X
43	flxA	251.29		
44	fliQ	251.26	X	X
45	flhE	231.94	X	
46	fliO	229.43	X	X
47	flhB	222.84	X	
48	flhA	221.21	X	X
49	ymdA	218.56		X
50	fliR	196.03	X	X
51	tsr	192.51		X
52	yibT	191.41		
53	yhjH	191.29		X
54	yjbJ	188.62		
55	hdeB	187.86		
56	slp	184.57		
57	ompF	181.05		
58	micF	178.49	negative	
59	gadE	178.32	negative	
60	hdeA	177.08		
61	hdeD	176.26		
62	gadX	173.90		
63	gadB	171.77		
64	gadA	165.48		
65	yjdA	148.57		
66	bssS	122.36		
67	ygiW	109.00		

^a Genes displayed here are sorted by the Log Bayes ratio (target gene versus non-target gene).

^b Blank indicates that the target gene shows the same pattern as the query gene.

Negative indicates that the target gene shows the inversed pattern as the query gene.

^c BEST identifies 29 genes among 46 target genes in RegulonDB. “X” indicates that the predicted gene is in the RegulonDB target set.

^d “X” indicates that the gene is predicted by CLR as a target gene.

Table S10. Information on the 56 potential FliA target genes identified by BEST in the 200-gene test set extracted from the *E. coli* compendium

Rank	Gene Name ^a	Log Bayes Ratio	positive/negative ^b	RegulonDB ^c	CLR ^d
1	fliZ	524.28		X	X
2	flgE	520.41		X	X
3	flgC	518.79		X	X
4	flgB	510.76		X	X
5	flgG	506.72		X	X
6	flgD	506.65		X	X
7	flgN	503.87			X
8	flgK	503.51			X
9	flgH	490.47		X	X
10	flgM	488.91			X
11	fliD	484.75		X	X
12	cheW	484.68		X	X
13	cheA	480.61		X	X
14	motB	470.69		X	X
15	motA	469.88		X	X
16	flgL	469.11		X	X
17	flgA	467.86			X
18	fliK	463.84		X	X
19	fliS	459.14		X	X
20	fliN	456.92		X	X
21	flgF	456.74		X	X
22	cheZ	453.19		X	X
23	cheR	452.47		X	X
24	fliL	452.27		X	X
25	fliJ	452.03		X	X
26	flgI	448.03		X	X
27	cheB	443.04		X	X
28	tar	439.80			X
29	fliC	439.60			X
30	fliF	435.47		X	X
31	fliM	431.91		X	X
32	fliG	429.41		X	X
33	cheY	425.70		X	X
34	flgJ	423.63		X	X
35	fliP	420.70		X	X
36	yecR	418.55			X
37	ycgR	415.26			X
38	tap	414.20		X	X
39	fliQ	404.20		X	X
40	fliH	389.79		X	X

41	flxA	377.76		X	X
42	fliT	369.77		X	X
43	ymdA	364.62			X
44	fliO	361.51		X	X
45	fliI	355.28		X	X
46	fliE	354.40		X	X
47	flhC	348.31			X
48	flhB	335.53		X	X
49	flhE	321.02		X	X
50	flhA	317.85		X	X
51	fliR	312.35		X	X
52	yhjH	310.92			X
53	flhD	305.93			X
54	tsr	281.05			X
55	ves	233.09			X
56	yjdA	187.37			X

^a Genes displayed here are sorted by the Log Bayes ratio (target gene versus non-target gene).

^b Blank indicates that the target gene shows the same pattern as the query gene.

Negative indicates that the target gene shows the inversed pattern as the query gene.

^c BEST indentifies 41 among 42 target genes in RegulonDB. “X” indicates that the predicted gene is in the RegulonDB target set.

^d “X” indicates that the gene is predicted by CLR as a target gene.

Figure S1. ROC curves for various query methods when applying to synthetic datasets simulated under different settings and when there are 100% foreground columns. BEST A default setting; BEST B allowing exclusion of individual cells from the foreground; BEST C fixing the indicator variables of five true target genes and five true experimental conditions as 1. **A.** No linear transformation nor cell-level noise. **B.** With linear transformation only. **C.** With cell-level noise only. **D.** With both linear transformation and cell-level noise.

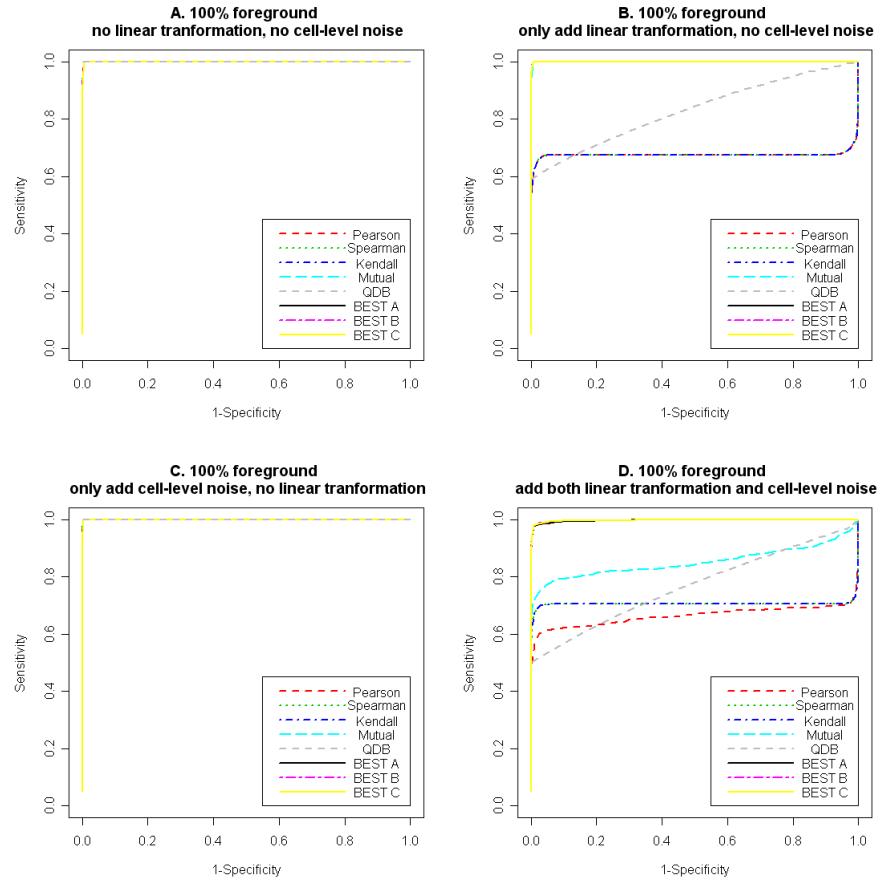


Figure S2. ROC curves for various query methods when applying to synthetic datasets simulated under different settings and when there are 75% foreground columns. BEST A default setting; BEST B allowing exclusion of individual cells from the foreground; BEST C fixing the indicator variables of five true target genes and five true experimental conditions as 1. **A.** No linear transformation nor cell-level noise. **B.** With linear transformation only. **C.** With cell-level noise only. **D.** With both linear transformation and cell-level noise.

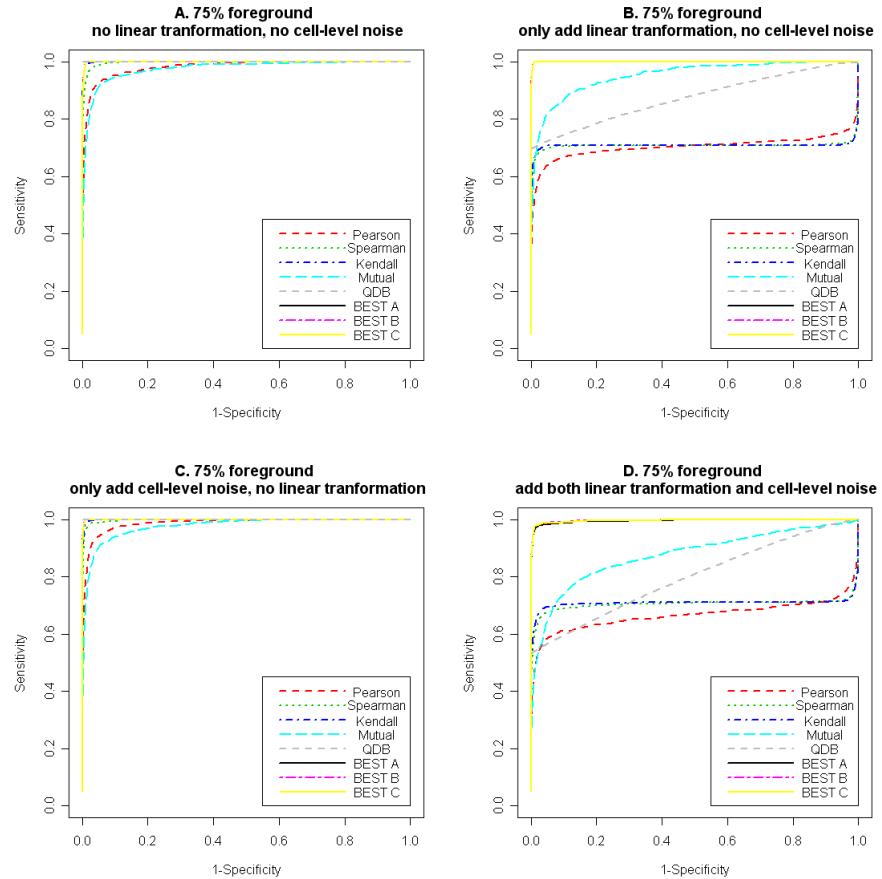


Figure S3. ROC curves for various query methods when applying to synthetic datasets simulated under different settings and when there are 50% foreground columns. BEST A default setting; BEST B allowing exclusion of individual cells from the foreground; BEST C fixing the indicator variables of five true target genes and five true experimental conditions as 1. **A.** No linear transformation nor cell-level noise. **B.** With linear transformation only. **C.** With cell-level noise only. **D.** With both linear transformation and cell-level noise.

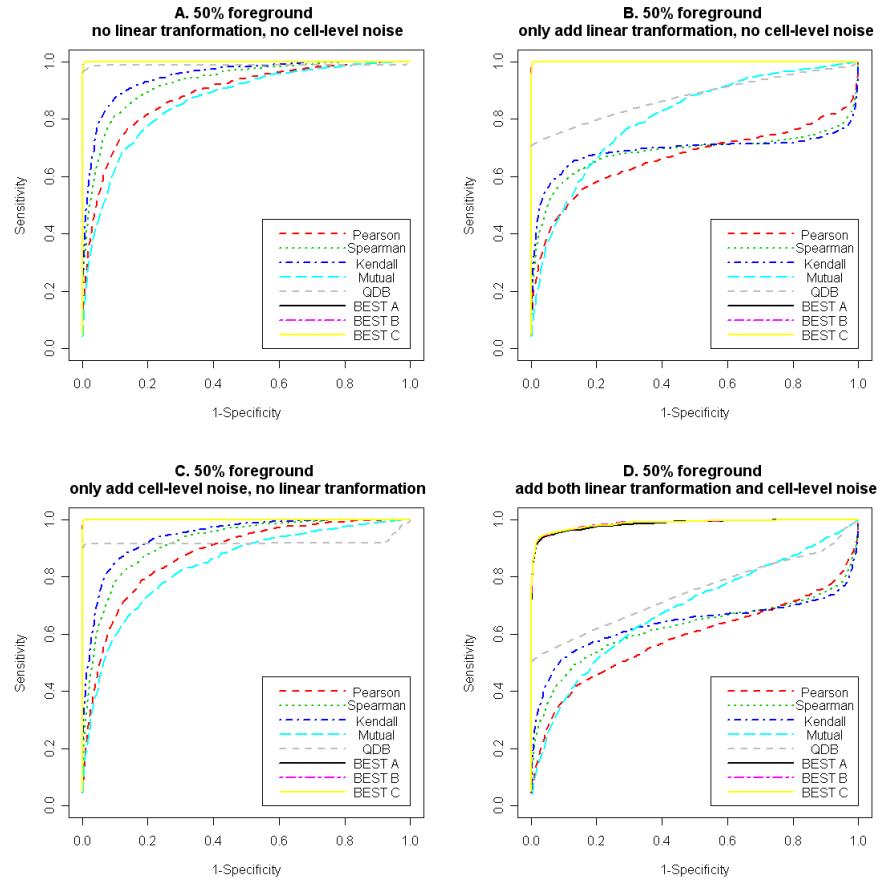


Figure S4. Log-likelihood trace plots of the ten parallel chains resulted from the BEST run on 100-gene and 300-gene test sets selected from the *E. coli* microarray compendium.

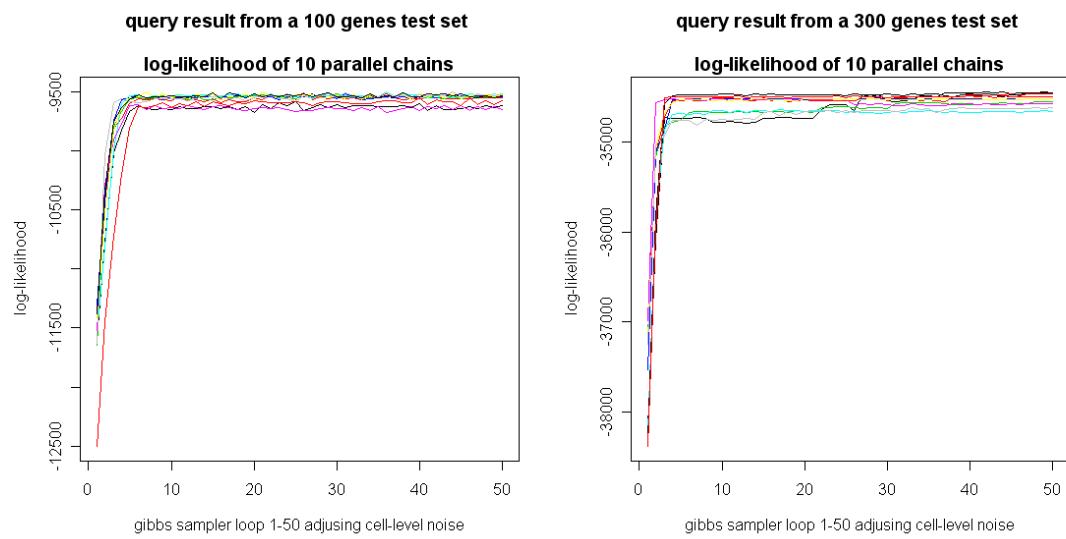
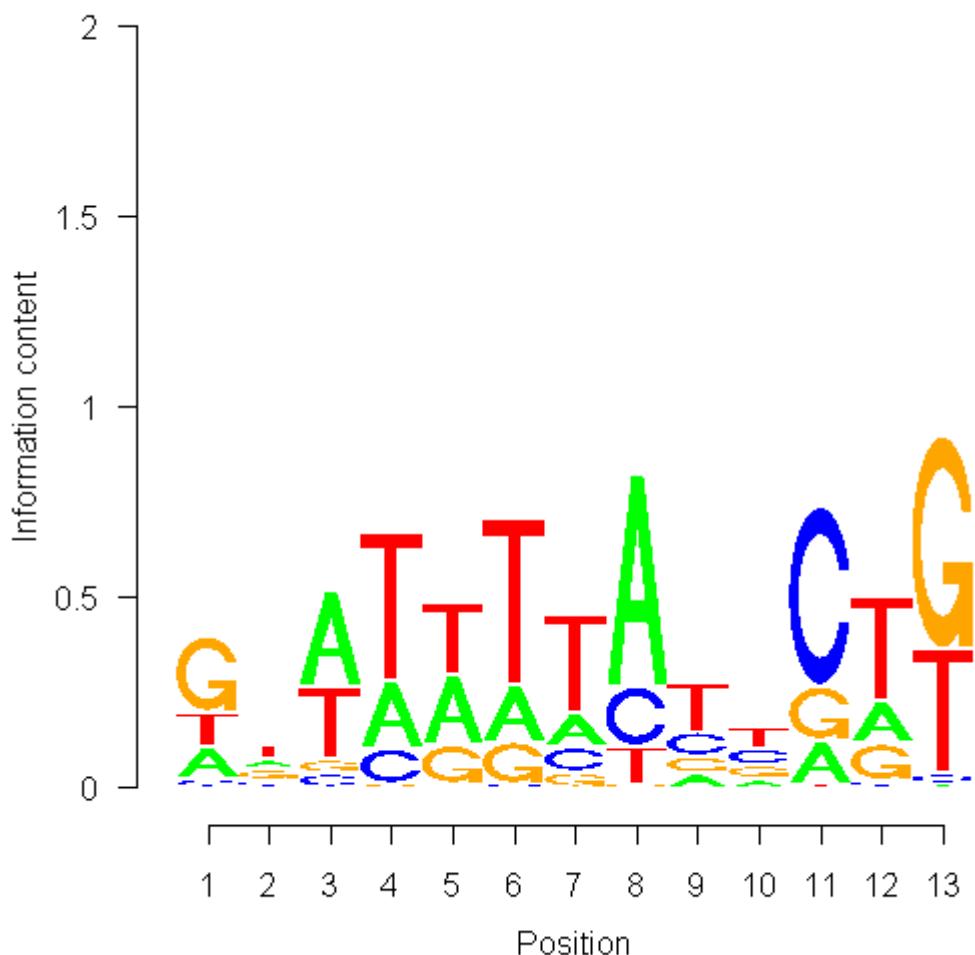


Figure S5. Sequence logo plot (Schneider and Stephens 1990) and position specific weight matrix (PSWM) for the motif of transcription factor Lrp. Lrp motif is downloaded from regulonDB:

http://regulondb.ccg.unam.mx/data/Matrix_AlignmentSet.txt. The logo plot was generated by the seqLogo program (Bembom 2007).



position	1	2	3	4	5	6	7	8	9	10	11	12	13
A	11	14	27	14	21	11	10	37	6	6	8	12	0
C	2	7	3	7	0	0	7	10	12	15	35	1	1
G	28	13	3	0	11	8	4	0	8	11	10	10	33
T	12	19	20	32	21	34	32	6	27	21	0	30	19

Figure S6. Boxplots of Pearson correlation coefficients. **A.** Boxplots of Pearson correlations between expression profiles of the 61 experimentally verified Lrp target genes and Lrp. The left one summarize correlations measured in the 162 background experiments and the right one summarize correlations measured in the 143 foreground experiments. A paired t-test comparing the two sets of correlation coefficients returns a p-value of 0.0079. **B.** Boxplots of Pearson correlations between expression profiles of the 28 genes BEST indentified as Lrp target. The left one summarize correlations measured in the 162 background experiments and the right one summarize correlations measured in the 143 foreground experiments. A paired t-test comparing the two sets of correlation coefficients returns a p-value of 1.948×10^{-12} .

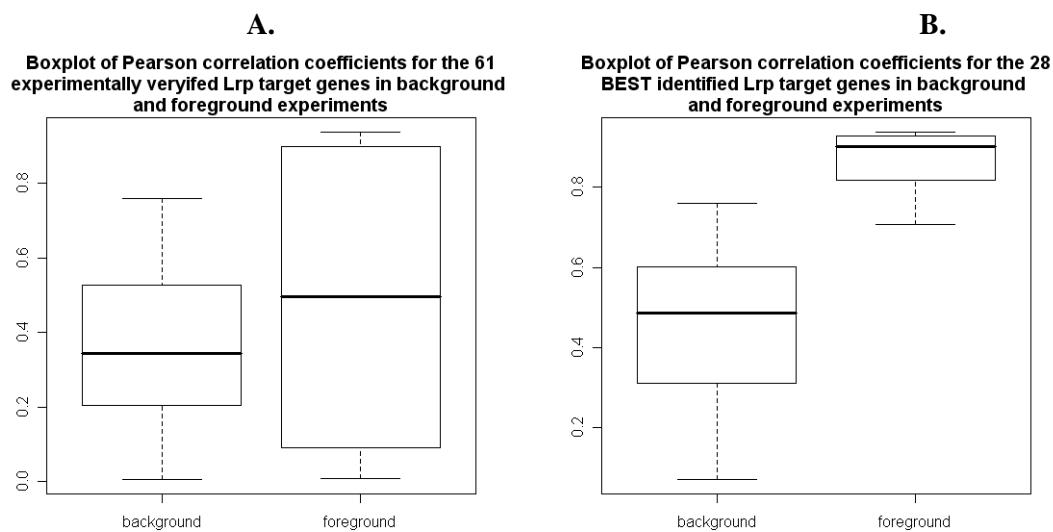


Figure S7. **A.** Histograms of expression profile differences (z_{ij}) in the top five experimental conditions (foreground). **B.** Histogram of expression profile differences (z_{ij}) in the bottom five experimental conditions (background). Data used here is the 100-gene test set selected from the *E. coli* microarray compendium.

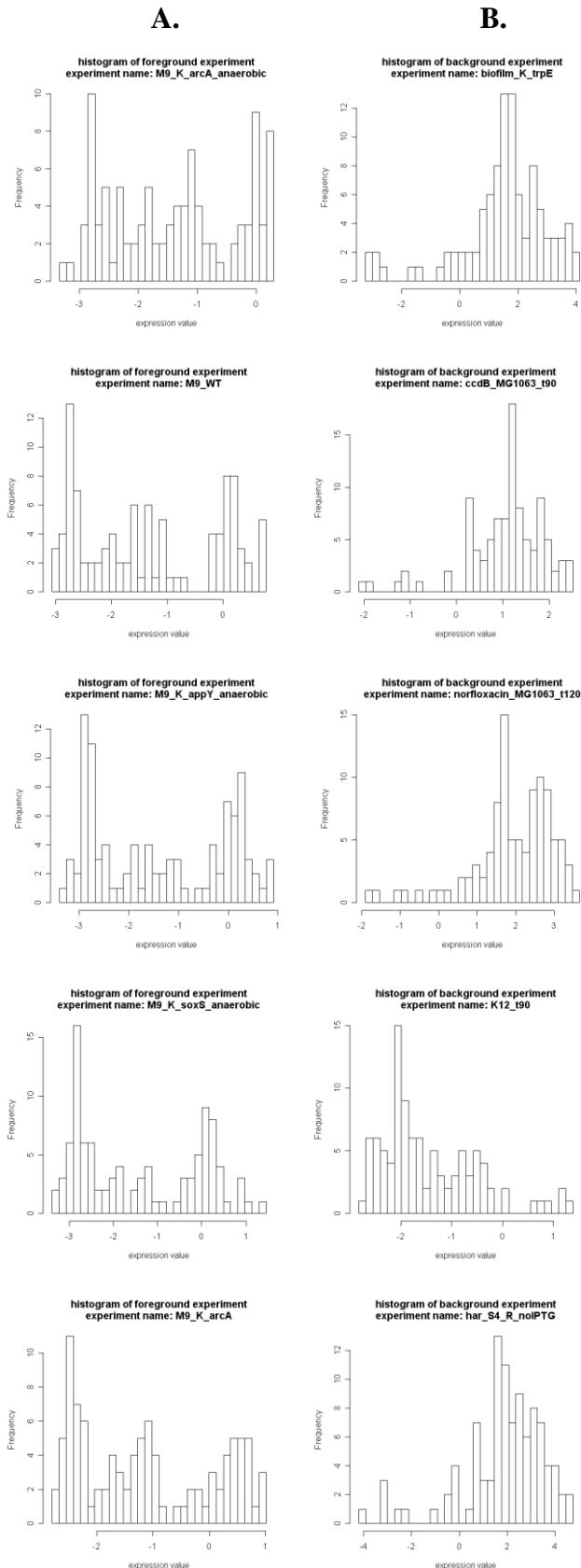
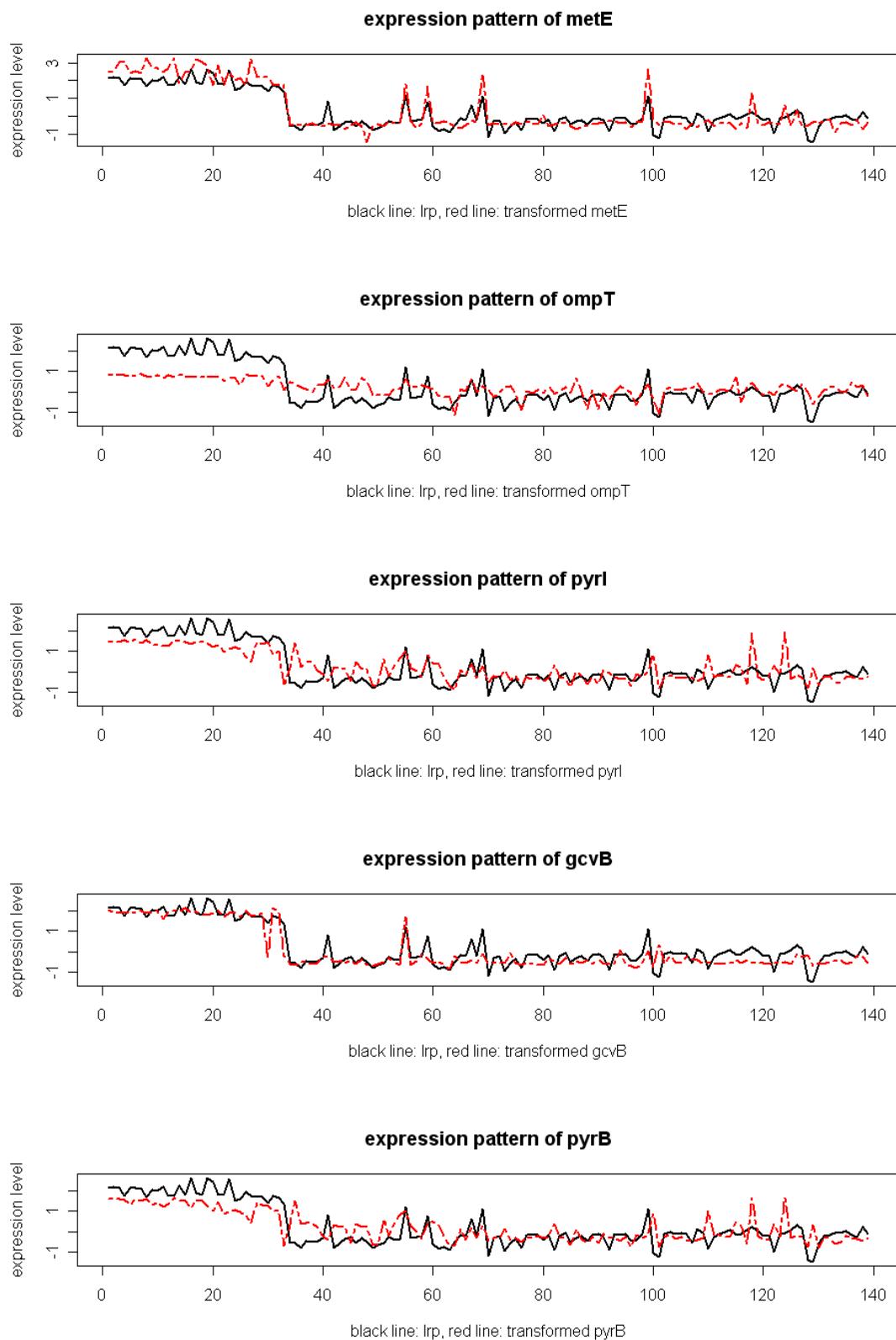
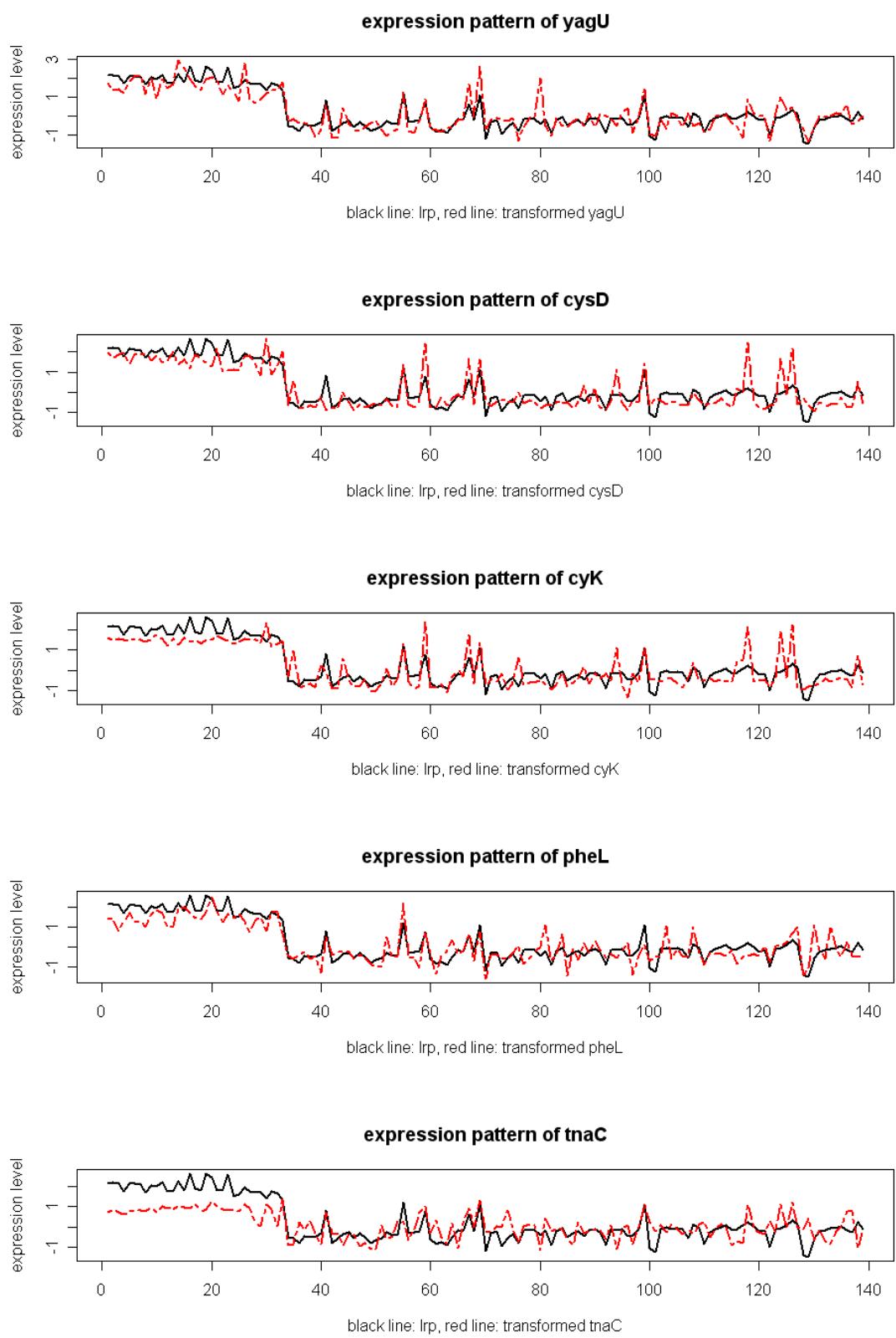
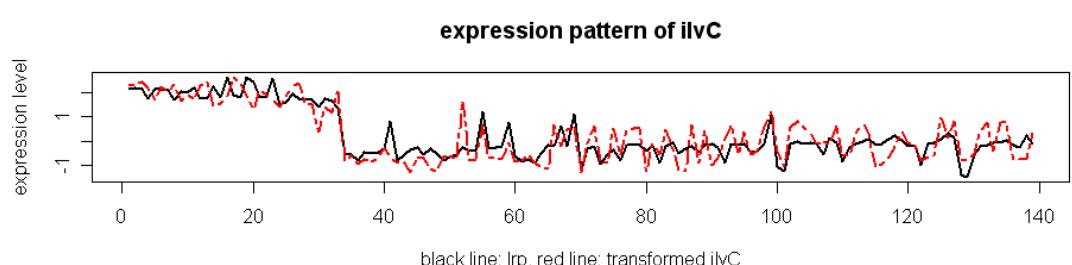
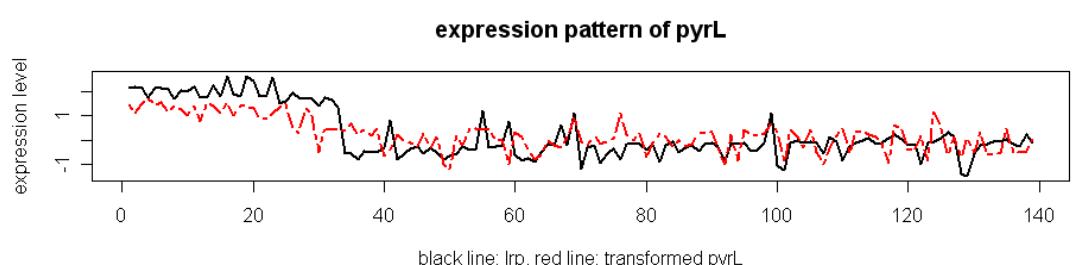
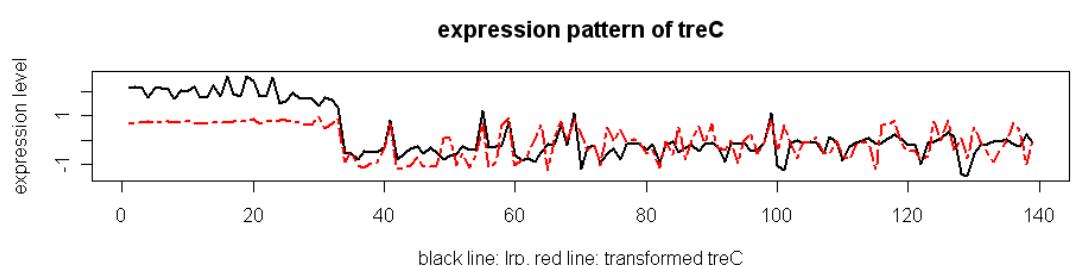
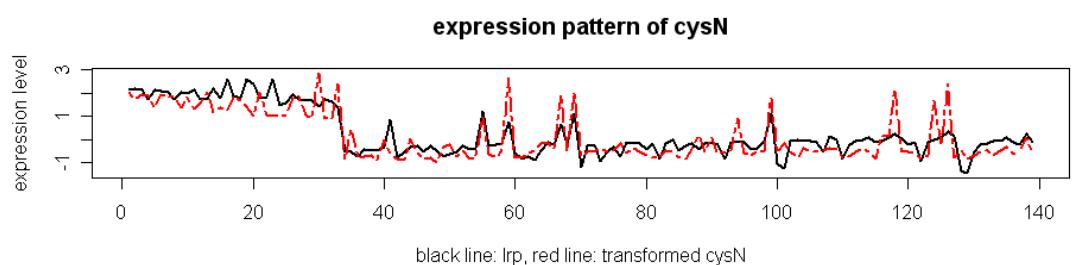
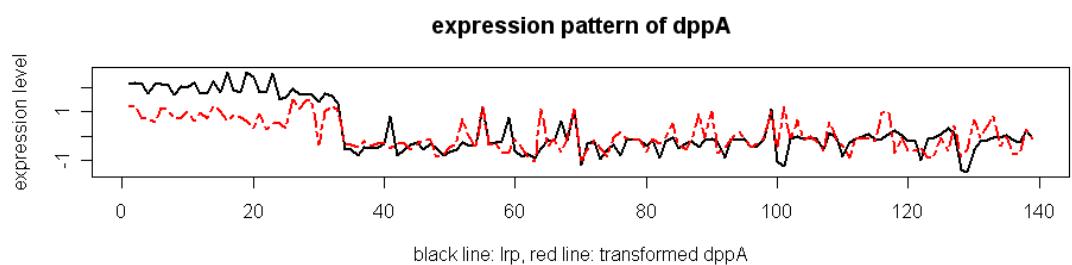
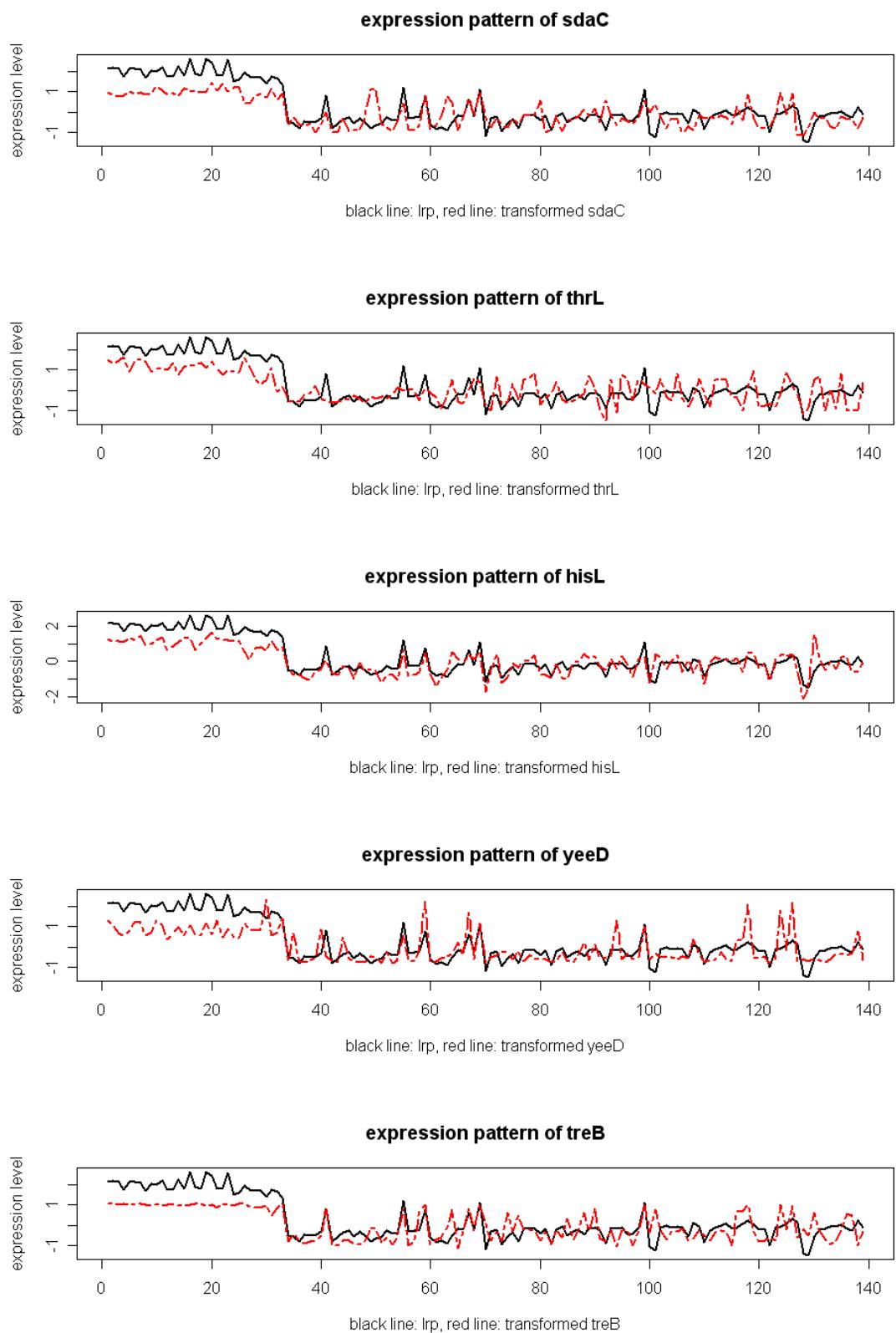


Figure S8. Trace plots of 24 predicted Lrp target genes identified by BEST that are not in the RegulonDB target set. Black lines indicate the query gene—Lrp, the red line indicate the potential target genes. Only the 139 foreground experimental conditions identified by BEST were shown in these plots.









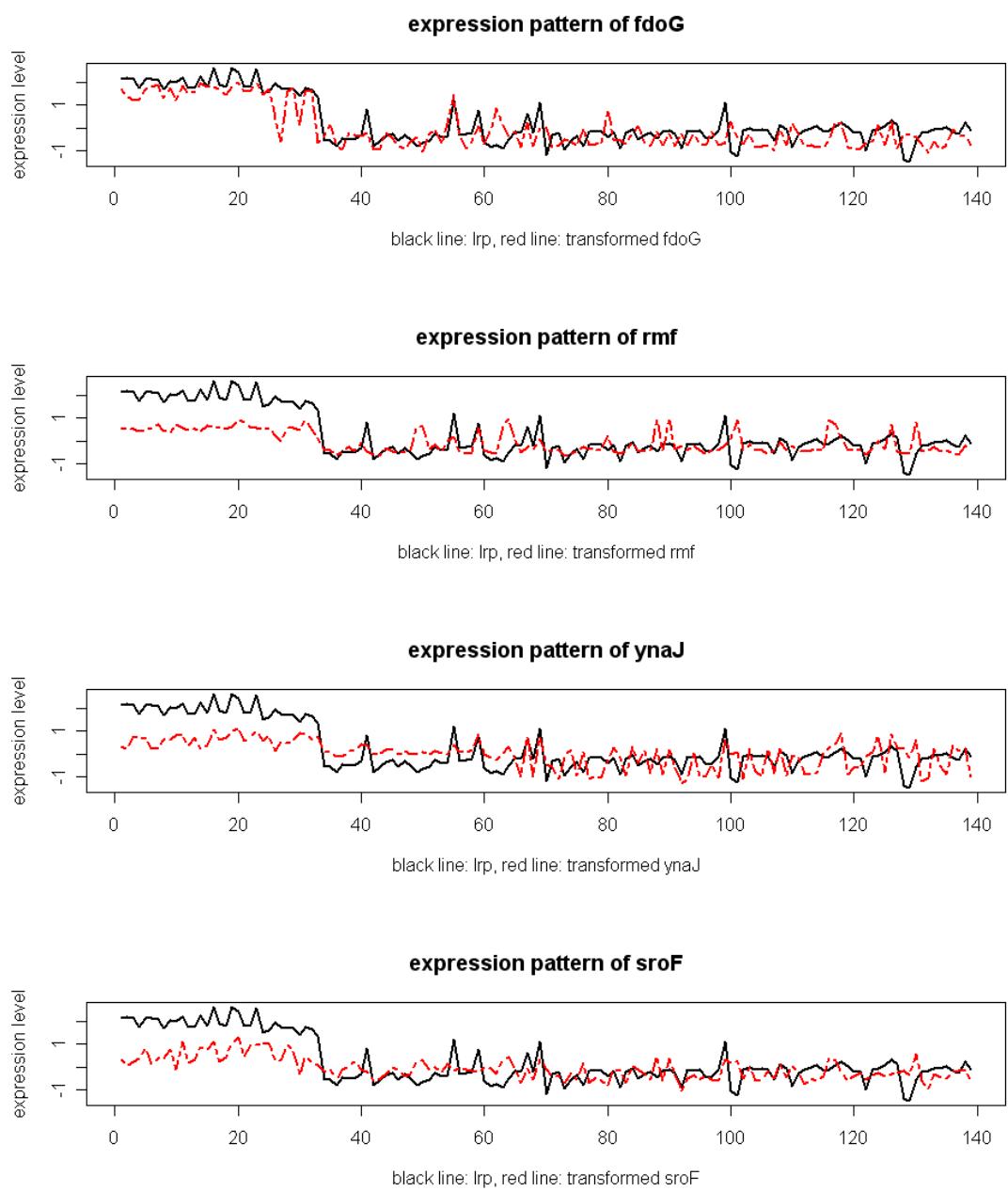
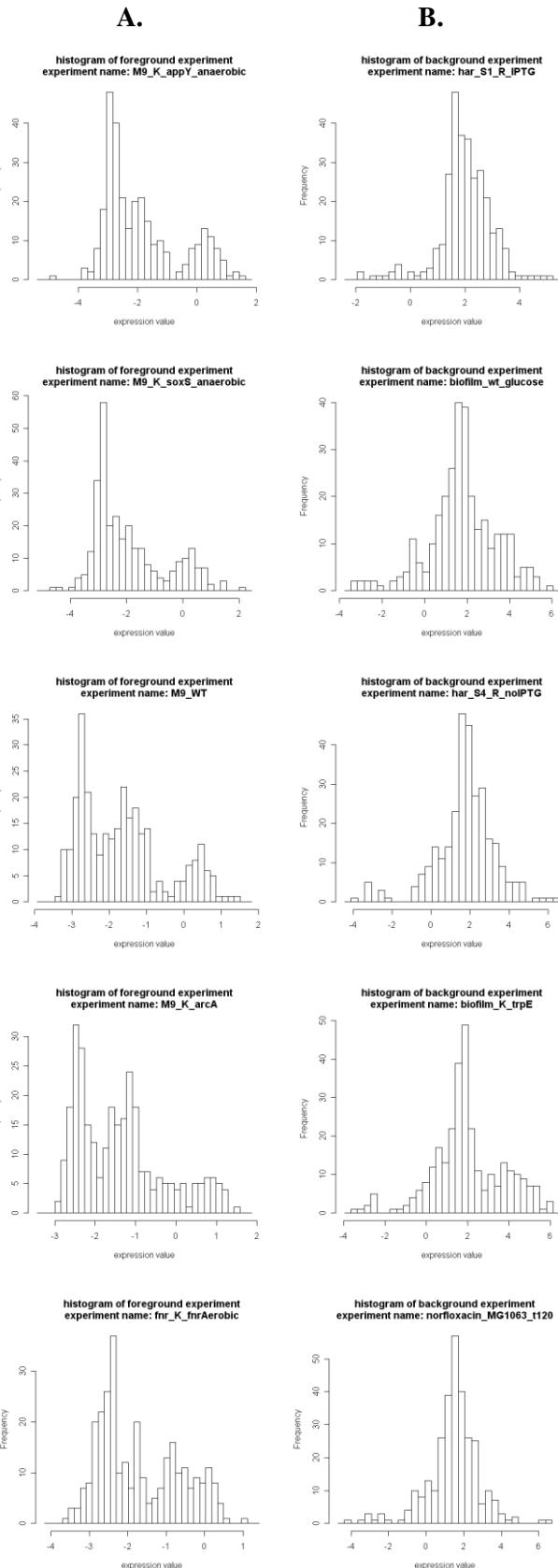


Figure S9. **A.** Histograms of expression profile differences (z_{ij}) in the top five experimental conditions (foreground). **B.** Histogram of expression profile differences (z_{ij}) in the bottom five experimental conditions (background). Data used here is the 300-gene test set selected from the *E. coli* microarray compendium.



Additional references:

1. Bembom O: **Sequence logos for DNA sequence alignments**. 2007.
2. Gelman A, Carlin JB, Stern HS, Rubin DB: **Bayesian data analysis**, Reprinted 1997. edn. London: Chapman & Hall; 1995.
3. Schneider TD, Stephens RM: **Sequence logos: a new way to display consensus sequences**. *Nucleic Acids Res* 1990, **18**(20):6097-6100.
4. Liu J: **Monte Carlo Strategies in scientific computing**. New York: Springer-Verlag; 2001.