Supplementary information – Notes, Figures, and Tables

Jan O. Korbel, Lars J. Jensen, Christian von Mering & Peer Bork:

Analysis of genomic context: Prediction of functional associations from conserved bidirectionally transcribed gene pairs

1. Species and clades used in the analysis

We used all 101 completely sequenced prokaryotic genomes available from the Proteome Analysis Database[1] downloaded on 3rd March 2003, excluding *Tropheryma whipplei* for which more than half of the genes lacked genomic coordinates. Genomes of the following species from 47 evolutionary clades (see **Supplementary Table 1**) were utilized (species names are the same as in STRING[2] version 4):

Aeropyrum pernix, Agrobacterium tumefaciens Cereon*, Agrobacterium tumefaciens* Wash*, Aquifex aeolicus, Archaeoglobus fulgidus, Bacillus halodurans, Bacillus subtilis, Bifidobacterium longum, Borrelia burgdorferi, Bradyrhizobium japonicum, Brucella melitensis, Brucella suis, Buchnera aphidicola* APS*, Buchnera aphidicola* Schiz*, Campylobacter jejuni, Caulobacter crescentus, Chlamydia muridarum, Chlamydia trachomatis, Chlamydophila pneumoniae* AR39*, Chlamydophila pneumoniae* CWL029*, Chlamydophila pneumoniae* J138*, Chlorobium tepidum, Clostridium acetobutylicum, Clostridium perfringens, Corynebacterium efficiens, Corynebacterium glutamicum, Deinococcus radiodurans, Escherichia coli* EDL933*, Escherichia coli* K12*, Escherichia coli* O157*, Escherichia coli* O6*, Fusobacterium nucleatum, Haemophilus influenzae, Halobacterium* sp. NRC-*1, Helicobacter pylori* 26695*, Helicobacter pylori* J99*, Lactococcus lactis, Leptospira interrogans, Listeria innocua, Listeria monocytogenes, Mesorhizobium loti, Methanococcus jannaschii, Methanopyrus kandleri, Methanosarcina acetivorans, Methanosarcina mazei, Methanothermobacter thermautotrophicum, Mycobacterium leprae, Mycobacterium tuberculosis* CDC1551*, Mycobacterium tuberculosis* H37Rv*, Mycoplasma genitalium, Mycoplasma pneumoniae, Mycoplasma pulmonis, Neisseria meningitidis* A*, Neisseria meningitidis* B*, Nostoc* sp. PCC 7120*, Oceanobacillus iheyensis, Pasteurella multocida, Pseudomonas aeruginosa, Pseudomonas putida, Pyrobaculum aerophilum, Pyrococcus abyssi, Pyrococcus furiosus, Pyrococcus horikoshii, Ralstonia solanacearum, Rickettsia conorii, Rickettsia prowazekii, Salmonella typhi* CT18*, Salmonella typhimurium* LT2*, Shewanella oneidensis, Shigella flexneri, Sinorhizobium meliloti, Staphylococcus aureus* MW2*, Staphylococcus aureus* Mu50*, Staphylococcus aureus* N315*, Staphylococcus epidermidis, Streptococcus agalactiae, Streptococcus mutans, Streptococcus pneumoniae* R6*, Streptococcus pneumoniae* TIGR4*, Streptococcus pyogenes M1, Streptococcus pyogenes* M18*, Streptococcus pyogenes* M3*, Streptomyces coelicolor, Sulfolobus solfataricus, Sulfolobus tokodaii,* *Synechococcus elongatus, Synechocystis* sp. PCC 6803*, Thermoanaerobacter tengcongensis, Thermoplasma acidophilum, Thermoplasma volcanium, Thermotoga maritima, Treponema pallidum, Ureaplasma parvum, Vibrio cholerae, Vibrio vulnificus, Wigglesworthia brevipalpis, Xanthomonas axonopodis, Xanthomonas campestris, Xylella fastidiosa, Yersinia pestis* CO92*, and Yersinia pestis* KIM.

Supplementary Table 1. Clades of evolutionarily closely related species obtained from the STRING server[2]. A significant subset of the 101 species considered is evolutionarily '*too closely*' related, i.e. large fractions of their genomes are still 'syntenic': shared gene arrangements are not indicative of evolutionary pressure, but simply of recent descent. Conservation of adjacent pairs was thus considered as relevant only if the pairs were conserved across distinct clades. We obtained evolutionary clades from STRING[2], which are manually refined groups of species on the basis of various phylogenetic distance measures, including conservation of gene arrangements. We tested other phylogenetic distance cutoffs (resulting in different clade compositions), always observing that co-directionally transcribed, or divergently transcribed, gene pairs were significantly more often conserved across species than convergently transcribed gene pairs (see also **Fig. 1**). Species currently lacking a closely related completely sequenced genome are forming 'single-species clades', and are not included in the table below.

Clades of closely related species
Helicobacter pylori J99*, Helicobacter pylori* 26695
Chlamydia muridarum, Chlamydia trachomatis, Chlamydophila pneumoniae J138*, Chlamydophila pneumoniae* CWL029*, Chlamydophila pneumoniae* AR39
Neisseria meningitidis A*, Neisseria meningitidis* B
Rickettsia prowazekii, Rickettsia conorii
Mycoplasma pneumoniae, Mycoplasma pulmonis, Mycoplasma genitalium, Ureaplasma parvum
Streptococcus pneumoniae R6*, Streptococcus pneumoniae* TIGR4*, Streptococcus pyogenes* M18*, Streptococcus pyogenes* M1*, Streptococcus mutans, Streptococcus pyogenes* M3*, Streptococcus agalactiae*
Pasteurella multocida, Haemophilus influenzae
Methanosarcina acetivorans, Methanosarcina mazei
Pyrococcus horikoshii, Pyrococcus abyssi, Pyrococcus furiosus
Thermoplasma volcanium, Thermoplasma acidophilum
Sulfolobus solfataricus, Sulfolobus tokodaii
Mycobacterium tuberculosis H37Rv*, Mycobacterium tuberculosis* CDC1551*, Mycobacterium leprae*
Brucella melitensis, Mesorhizobium loti, Sinorhizobium meliloti, Agrobacterium tumefaciens Cereon, Agrobacterium tumefaciens Wash, Brucella suis, Bradyrhizobium japonicum
Xanthomonas axonopodis, Xanthomonas campestris, Xylella fastidiosa
Vibrio cholerae, Yersinia pestis CO92*, Salmonella typhimurium* LT2*, Salmonella typhi* CT18*, Escherichia coli* K12*, Escherichia coli* O157*, Escherichia coli* EDL933*, Shewanella oneidensis, Vibrio vulnificus, Yersinia pestis* KIM*, Shigella flexneri, Escherichia coli* O6
Buchnera aphidicola APS*, Buchnera aphidicola Schiz, Wigglesworthia brevipalpis*
Clostridium perfringens, Clostridium acetobutylicum
Listeria innocua, Listeria monocytogenes, Staphylococcus aureus MW2*, Staphylococcus aureus* N315*, Staphylococcus aureus* Mu50*, Bacillus halodurans, Bacillus subtilis, Oceanobacillus iheyensis, Staphylococcus epidermidis*
Corynebacterium efficiens, Corynebacterium glutamicum
Pseudomonas aeruginosa, Pseudomonas putida

2. Orthology transfer and benchmarking of functional associations

Orthology-transfer was used to predict functional associations between *E. coli* genes, which are bidirectionally transcribed only in other species. Transfer of orthology was performed using orthologous groups from STRING[2] version 4 (clusters originally obtained from the COG database[3] and subsequently expanded[2] to cover all completely sequenced genomes available). In cases where the transfer was ambiguous, i.e. when more than one *E. coli* protein was assigned to an orthologous group, we selected the *E. coli* protein representing the '*best hit*'. (*Best hits* were obtained using Smith-Waterman searches against the *E. coli* proteome. For each *E. coli* protein, we averaged the bit scores of hits from all *DT-pairs* of the genomes in question.) Transferred associations were added to functional links derived from conserved *DT-pairs* present in the *E. coli* genome. (When performing benchmarking, more than 50% of the final associations were due to orthology-transfer.) Pairs with '*XX*' classification were benchmarked using a framework reported earlier[2], applying KEGG maps[4] at the level of protein coding genes: predicted associations mapping to *E. coli* genes assigned to the same map were taken as 'true positive' predictions. Since KEGG maps often lack the associated regulators, we benchmarked *DT-pairs* classified as '*RX*' using annotated transcription regulatory interactions[5-7], considering a predicted functional link as 'true' if the respective regulator can be associated with a putative target gene or process using regulatory interactions or KEGG maps. Co-directionally transcribed gene pairs and gene fusions were benchmarked using the same procedure as for pairs with '*XX*' classification.

A comparison of our novel method with previous approaches that exploit the genomic context of genes for function prediction is shown in the Box "Comparison of genomic context approaches". Previous approaches were applied using the STRING[2] server at an accuracy-level of 40%. For the *DT-pair* method, we considered all pairs conserved across at least 3 clades, which corresponds to a roughly equivalent accuracy.

Benchmarking functional associations for well-resolved orthologous groups

Accuracy values estimated for '*RX*' pairs (as described above) represent a lower estimate of the actual picture, due to an enrichment of transcriptional regulators in poorly resolved orthologous groups[8] containing two or more similar regulators per species, which complicates transfer of function via orthology. If only well-resolved orthologous groups were considered in the benchmark (i.e. using an inparalog-corrected genes-species ratio of 4; ref. 8), we found an accuracy of 75% for *DT-pairs* with '*RX*' classification conserved across three or more clades, while we estimated 71% for co-directional pairs and 81% for gene fusions (for both of the latter, we did not separate '*XX*' and '*RX*' classification) using the given parameters. This indicates that given well-resolved orthology relationships, our predictor allows precise statements of type and nature of the predicted association.

Benchmarking functional associations of genes predicted from conserved bidirectionally transcribed operons

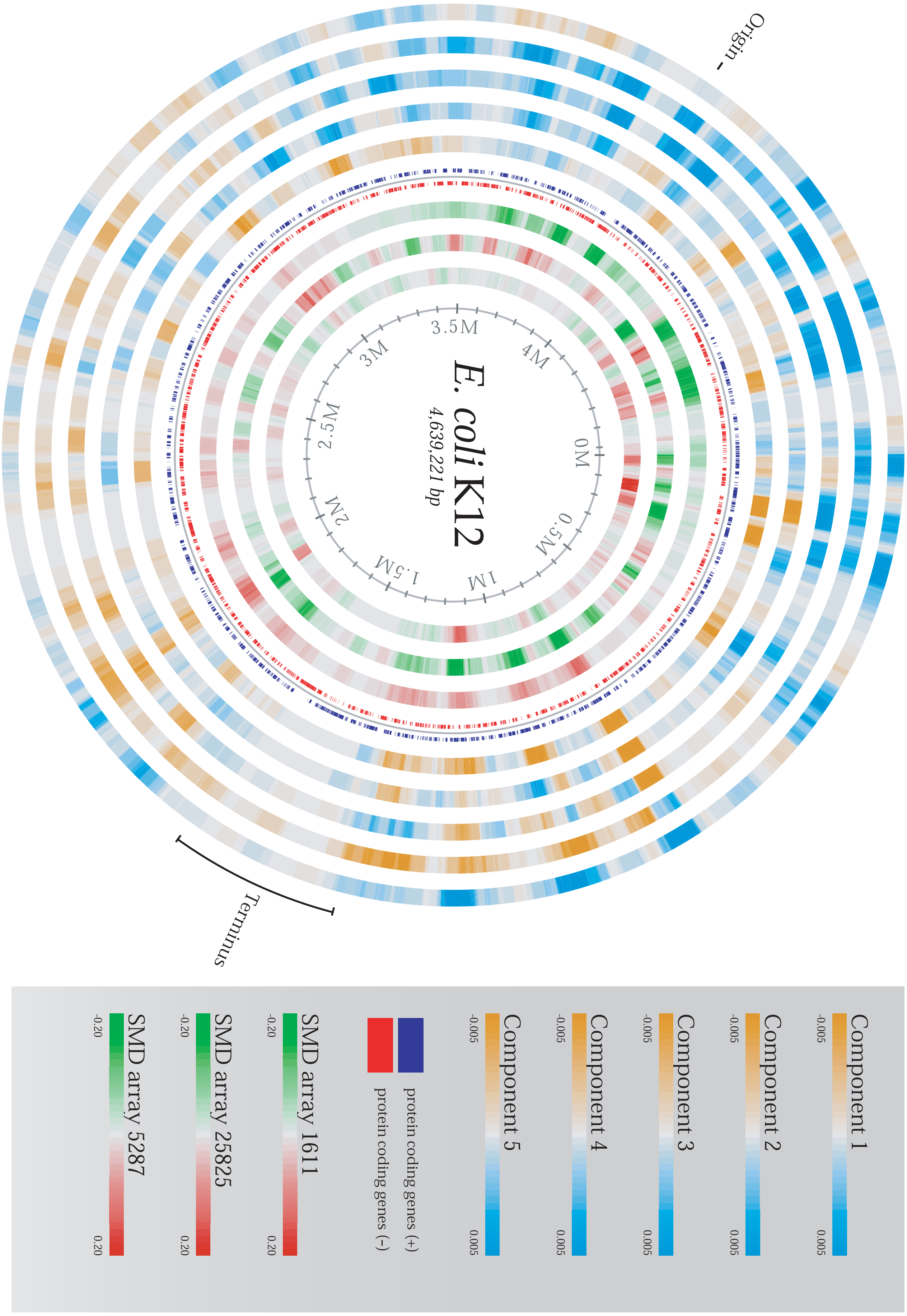
Often, a conserved operon structure extends from at least one side of a conserved *DT-pair* as illustrated in the Box “Comparison of genomic context approaches”. Since conserved operons strongly indicate functional association, we alternatively predicted functional relationships between genes sitting in conserved bidirectionally transcribed operon structures: i.e. we predicted pairwise functional associations between all genes transcribed in opposite directions (using the same definition for putative operons as in STRING; i.e. adjacent co-directionally transcribed genes with an intergenic distance of 300 bp or less are assumed to be part of the same operon). Such an analysis revealed 1810 pairwise associations between orthologous groups[2, 9] for pairs conserved across at least 3 clades (7630 are conserved across 2 clades or more). Using this strategy, the accuracy for ‘*RX*‘ pairs (representing the majority of extracted pairs) was 38% (for pairs conserved across at least 3 clades). When only considering well-resolved orthologous groups (inparalog-corrected genes-species ratio of 4; see above), we found an accuracy of 59% for ‘*RX*’ pairs.

All pairs conserved across at least 3 clades were used for the comparison with previous genomic context-based methods (see Box “Comparison of genomic context approaches”).

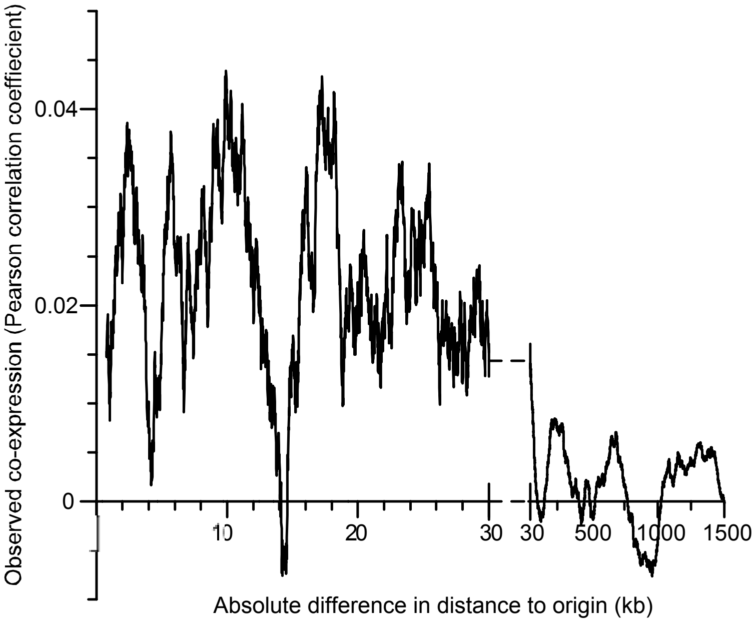
3. The observed long-range correlations in expression go beyond effects caused by relative distance to the origin of replication

In order to quantify associations between nearby genes globally, we mapped gene expression data from 95 previously published DNA microarray experiments onto the genome sequence of *E. coli* K12. Surprisingly, even genes more than 100 kb apart appear significantly co-expressed indicating coordinative regulation (see **Supplementary Fig. 1**, or **Box 2**), an effect going far beyond the expected associations via operons (the average length of transcriptional units in the *E. coli* genome is estimated to be in the order of 1.7 kb; ref. 10). In order to test whether the effect also goes beyond the known correlation in expression depending on the distance relative to the origin of replication[11], we undertook the following analysis: we divided the circular chromosome of *E. coli* into two equal halves, cutting at the origin and the terminus. Then we compared the expression of genes positioned on distinct halves, plotting correlation in expression versus the absolute difference in distance of the genes to the origin (**Supplementary Fig. 2**). Resulting average expression correlation coefficients (Pearson correlation) are around 0.02-0.025, dropping to a value close to 0 after only 80 kb. Thus, the long-range correlations in expression we observed go beyond effects caused by relative distance to the origin of replication.

Supplementary Figure 1



Supplementary Figure 1. Long range expression correlations in the *E. coli* genome go beyond the expected expression correlation of genes co-localized in operons (*see previous page*). The correlation in expression between nearby genes was measured using gene expression data from 95 previously published DNA microarray experiments[12], and visualized using the GenomeAtlas software[13,14]. After normalization[15] the log-ratios for all arrays were combined into a matrix (assigning a log-ratio of zero in the case of missing values). Principal component analysis was performed on this matrix to reduce its dimensionality while retaining as much of the variance as possible (e.g. this allows visualizing 26% of the variance in the first 5 principal components, roughly twice as much as observed when shuffling the genes within each array). The first 5 principal components (outer circles), along with 3 representative arrays (inner circles), are shown as a GenomeAtlas[13,14] by mapping expression data for each gene onto its corresponding genomic position, applying a running average of 50 kb. Each circle visualizes a separate principal component or array using a color scale that highlights genomic regions where genes are preferentially up- or down-regulated (yellow/blue or green/red, respectively). Thin circles (namely the 6th, dark blue, and the 7th, dark red, circle counting from outside) indicate positions of protein coding genes on both strands of the chromosome. Component 2 divides the genome into two parts perpendicular to an axis defined by the origin of replication, and the terminus region – thus capturing at least in part the known correlation in expression depending on the distance relative to the origin of replication[11]. In all components, large regions of similarly expressed genes were observed.



Supplementary Figure 2. Weak correlations in expression of *E. coli* genes caused by similar distance relative to the origin of replication. We divided the circular *E. coli* genome into two equal halves cutting at origin and terminus, and determined the average expression correlation of genes from distinct halves. Namely, we plotted the correlation in expression of two genes positioned on different sides of the genome versus the absolute difference in distance to the origin of replication. When relative distances to the origin are comparable, genes are indeed correlated in expression; however, the average expression correlation is around 0.02–0.025 (Pearson correlation coefficient), dropping to ~0 after only 80 kb – thus markedly below the long-range expression correlations we observed in *E. coli* (see Box “Genomic vicinity and gene co-expression in prokaryotes”).

4. Adjacent bidirectionally transcribed genes with conserved organization are significantly co-expressed

We use the definition '*conserved gene pair*' for a pair of neighboring protein-coding genes that have corresponding adjacent orthologs with conserved gene orientation in a genome from an evolutionarily distant clade (see **Supplementary Table 1** for clade assignments). More widely conserved pairs have corresponding orthologs in more than one clade. We ignored adjacent genes with overlapping coding regions, since DNA microarrays used in the expression analysis were not strand dependent. We moreover ignored adjacent paralogous genes which are members of a single orthologous group[2, 3], as such pairs of genes likely originated from recent duplications rather than from evolutionary conservation.

In order to quantify associations between adjacent divergently transcribed genes, we analyzed *E. coli* microarray expression data and found that *DT-pairs* – if their orientation has been conserved in at least one distant clade – appear significantly more often co-expressed than non-conserved *DT-pairs*, or than genes randomly picked from the genome (both at the 0.001-level; Kolmogorov-Smirnov, or KS test; see **Fig. 3** and **Supplementary Table 2**). Going to even higher levels of conservation, namely analyzing *DT-pairs* conserved in several independent clades, significantly increases the co-expression as compared to pairs conserved in only one distant clade ($P<0.05$, KS test).

When analyzing conserved *convergently* transcribed gene pairs, we found that the mean expression correlation (Pearson correlation coefficient τ) of conserved and non-conserved pairs differs by less than 1% – a negligible difference given that only 24 conserved gene pairs were found in *E. coli*. Despite the small number of pairs, conserved *DT-pairs* are significantly more often co-expressed than convergently transcribed pairs ($P<0.05$, KS test). While the mean expression correlation coefficients differ considerably (τ=0.43 for the former, and τ=0.26 for the latter), the significance value is marginal due to the small set of conserved convergently transcribed genes.

Surprisingly, the co-expression of widely conserved *DT-pairs* is comparable to co-directionally transcribed gene pairs, likely corresponding to genes in operons: we determined an average expression correlation of τ=0.53 for highly conserved *DT-pairs*, and τ=0.66 for highly conserved co-directional pairs, while co-directionally transcribed *E. coli* pairs in general have a Pearson correlation coefficient of τ=0.52.

Supplementary Table 2. *E. coli DT-pairs* with conserved gene organization are more often co-expressed than non-conserved *DT-pairs*, or than randomly selected genes. Expression correlation is indicated in terms of numbers of genes (fractions given in brackets) having an expression correlation higher than given Pearson correlation coefficient cut-offs. Values above 0 indicate co-expression and hence co-regulation, and values above 0.6 indicate functional association[16]. Black numbers indicate pairs of genes randomly picked from the genome. Red values were generated using all 95 DNA microarrays from SMD[12] (this data contains the least noise: considering a large variety of experimental conditions allows averaging out 'accidental' expression of genes). Blue values were produced limiting the analysis to the 50 microarrays submitted to SMD by Arkady Khodurski (University of Minnesota) and coworkers, thus relying on data from just one experimental laboratory. Data in green was generated limiting the analysis to replicated experiments (19 microarrays).

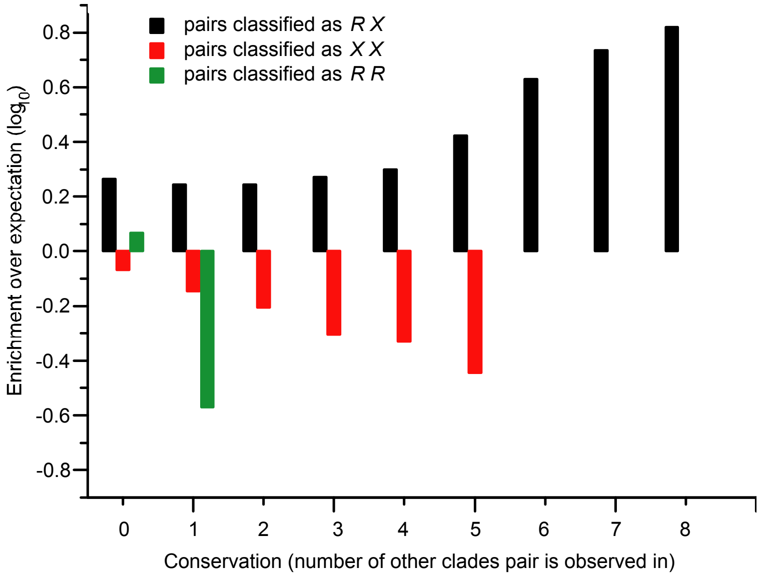
	total number of available pairs	***pairs with expression correlation coeff. >0.6***	***pairs with expression correlation coeff. >0***
Randomly selected gene pairs	16517	631 (3.82%)	8223 (49.8%)
Divergently transcribed neighbouring gene pairs	572 533 533	133 (23.3%) 105 (19.7%) 104 (19.5%)	482 (84.3%) 396 (74.3%) 373 (70.0%)
Divergently transcribed, orientation conserved in distant clades	148 148 148	55 (37.2%) 41 (27.7%) 43 (29.1%)	129 (87.2%) 115 (77.7%) 110 (74.3%)
Divergently transcribed, orientation conserved in several clades	58 58 58	27 (46.6%) 23 (39.7%) 20 (34.5%)	55 (94.9%) 48 (82.8%) 45 (77.6%)

5. Conserved *DT-pairs* are enriched in genes encoding transcriptional regulators

Determining which genes are typically involved in conserved *DT-pairs*, we observed a strong enrichment of pairs classified as '*RX'*, in which one of the genes encodes a transcriptional regulator (*'R'*), and the other a non-regulatory protein (*'X'*) (see **Supplementary Table 3** and **Supplementary Fig. 3**).

Supplementary Table 3. '*RX'* pairs, in which one of the genes encodes a transcriptional regulator, are highly enriched among conserved *DT-pairs* – while mostly pairs with '*XX'* classification were observed among conserved *co-directionally* transcribed genes. Shown are cumulative counts for *DT-pairs* (←→), and co-directional pairs (→→); numbers in brackets indicate observed fractions of '*RX',* *'XX',* and '*RR'* pairs. (Here, we ignored pairs in which at least one gene is member of a non-supervised orthologous group[8] (NOG), or of an orthologous group obtained from the COG database[15] (COG), which has been annotated as '*Uncharacterized*'.)

	conservation (number of other clades pair is observed in)									
←→	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
RX	5268 (0.27)	2428 (0.51)	1501 (0.63)	1133 (0.71)	738 (0.71)	572 (0.74)	442 (0.97)	387 (1.0)	152 (1.0)	82 (1.0)
XX	14125 (0.72)	2335 (0.49)	868 (0.37)	456 (0.29)	303 (0.29)	196 (0.26)	15 (0.03)	- (-)	- (-)	- (-)
RR	159 (0.008)	41 (0.009)	4 (0.002)	- (-)	- (-)	- (-)	- (-)	- (-)	- (-)	- (-)
→→										
RX	10071 (0.12)	4258 (0.10)	2618 (0.08)	1908 (0.07)	1534 (0.06)	1217 (0.06)	998 (0.05)	891 (0.05)	813 (0.05)	759 (0.05)
XX	74111 (0.88)	38877 (0.90)	30715 (0.92)	26510 (0.93)	22989 (0.94)	20664 (0.94)	18690 (0.95)	17292 (0.95)	15808 (0.95)	14160 (0.95)
RR	464 (0.005)	237 (0.005)	152 (0.005)	89 (0.003)	59 (0.002)	44 (0.002)	35 (0.002)	35 (0.002)	35 (0.002)	35 (0.002)



Supplementary Figure 3. Enrichment of *DT-pairs* classified as '*RX*'. Fractions of '*RX*', '*XX*' and '*RR*' pairs (based on cumulative counts) were compared to expected values: i.e. we shuffled 101 prokaryotic genomes 6,000 times to obtain expected numbers of pairs for different levels of "*conservation*" across clades. Bars are shown for all cases where observed *DT-pairs* stem from >5 distinct pairs of orthologous groups. The log-linear plot reveals 1.8-6.5 enrichment of pairs classified as '*RX*' (depending on the level of conservation), while '*XX*' pairs and '*RR*' pairs are significantly under-represented. For '*RX*' pairs, the initial slight decrease of enrichment is due to the fact that several transcriptional regulators are assigned in poorly resolved orthologous groups[2,3,8]; i.e. they are members of large protein families with several paralogous proteins per genome. Although the fraction of *DT-pairs* classified as '*RX*' increases strongly with conservation (see **Supplementary Table 3**), poorly resolved orthologous groups are more likely to form adjacent pairs at random.

6. *E. coli* transcriptional regulators encoded in conserved *DT-pairs* mostly regulate the divergently transcribed gene, as well as their own biosynthesis

To test whether transcriptional regulators residing in conserved *DT-pairs* correctly predict regulatory interactions, we analyzed all 135 regulators of *E. coli* having at least one known, annotated target gene[5-7], and found that regulators encoded in *DT-pairs* with evolutionarily conserved gene orientation mostly regulate the divergently transcribed gene, as well as their own biosynthesis in an auto-regulatory manner (**Supplementary Table 4**).

Supplementary Table 4. *E. coli* transcriptional regulators encoded in conserved *DT-pairs* are likely to regulate the divergently transcribed gene, as well as their own biosynthesis via auto-regulation. Here we considered all 22 transcriptional regulators of *E. coli* residing within conserved *DT-pairs*, which have at least one known and annotated target gene[5-7]. For each pair, we show the degree of conservation, and given evidence for regulation of the divergently transcribed gene, as well as for auto-regulation.

***E. coli* transcriptional regulator (Swissprot ID)**	**Divergently transcribed gene**	**Other clades pair has been conserved in**	**Evidence for regulation of divergent promoter**	**Evidence for auto-regulation**
BetI (BETI_ECOLI)	*betT* (BETT_ECOLI)	1	yes[5]	yes[5]
CynR (CYNR_ECOLI)	*cynT* (CYNT_ECOLI)	1	yes[5]	yes[5]
Lrp (LRP_ECOLI)	*trxB* (TRXB_ECOLI)	1	*regulator unknown*	yes[5]
FadR (FADR_ECOLI)	*nhaB* (NHAB_ECOLI)	1	*regulator unknown*	*unknown*
LrhA (LRHA_ECOLI)	*yfbQ* (YFBQ_ECOLI)	1	*regulator unknown*	yes[5]
GcvR (GCVR_ECOLI)	*dapA* (DAPA_ECOLI)	2	*regulator unknown*	*no auto-regulation*[17]
HcaR (HCAR_ECOLI)	*hcaE* (HCAE_ECOLI)	3	yes[7]	yes[7]
LysR (LYSR_ECOLI)	*lysA* (DCDA_ECOLI)	2	yes[5]	yes[5]
GlcC (GLCC_ECOLI)	*glcD* (GLCD_ECOLI)	1	yes[5]	yes[5]
ArgR (ARGR_ECOLI)	*mdh* (MDH_ECOLI)	1	no; other regulators exist[5]	yes[5]
Crp (CRP_ECOLI)	*yhfA* (YHFA_ECOLI)	1	yes[5]	yes[5]
YiaJ (YIAJ_ECOLI)	*yiaK* (YIAK_ECOLI)	1	yes[5]	yes[18]
DnaA (DNAA_ECOLI)	*rmpH* (RL34_ECOLI)	7	*regulator unknown*	yes[5]
AsnC (ASNC_ECOLI)	*asnA* (ASNA_ECOLI)	1	yes[5]	yes[5]
IlvY (ILVY_ECOLI)	*ilvC* (ILVC_ECOLI)	1	yes[5]	yes[5]
MetR (METR_ECOLI)	*metE* (METE_ECOLI)	3	yes[19]	yes[7]
GlnA (GLNA_ECOLI)	*typA* (TYPA_ECOLI)	1	*regulator unknown*	yes[7]
SoxS (SOXS_ECOLI)	*soxR* (SOXR_ECOLI)	1	*regulator unknown*	yes[7]
SoxR (SOXR_ECOLI)	*soxS* (SOXS_ECOLI)	1	yes[5]	yes[5]
MelR (MELR_ECOLI)	*agaL* (AGAL_ECOLI)	2	yes[5]	yes[5]
AcrR (ACRR_ECOLI)	*acrA* (ACRA_ECOLI)	7	yes[7]	evidence for autoregulation from *Enterobacter aerogenes*, close relative of *E. coli* with divergently transcribed *acrR* and *acrA*[20]
RpoE (RPOE_ECOLI)	*nadB* (NADB_ECOLI)	1	recent study indicates the co-regulation of *rpoE* and *nadB* in *Yersinia enterolytica*[21], close relative of *E. coli*	yes[22, 23]

7. Proteins encoded by *DT-pairs* classified as '*XX*' may act as post-transcriptional regulators

Notably, a considerable fraction of characterized proteins encoded by *DT-pairs* classified as '*XX*' may in fact act as post-transcriptional regulators: we found evidence for several of such cases among all *E. coli DT-pairs* conserved across at least four clades, which were classified as '*XX*' (see **Supplementary Table 5**). In order to assess whether the observed fraction of nucleic acid binding proteins (a considerable fraction of which might act in transcriptional or post-transcriptional gene expression regulation) is above expectation, we analyzed functional categories of randomly selected *E. coli* genes with similar phylogenetic coverage and inparalog-corrected genes-species ratios[8]. We found that the fraction of known nucleic acid binding proteins is indeed more than two-fold higher than expected for widely conserved *E. coli DT-pairs* with '*XX*' classification.

Supplementary Table 5. Some proteins encoded by *DT-pairs* classified as '*XX*' likely act as post-transcriptional regulators. Shown are all *E. coli DT-pairs* conserved across at least four clades (i.e. besides in *E. coli* the pair is present in three additional clades), which have '*XX*' classification. (We excluded pairs for which at least one of the corresponding orthologous groups[3] was annotated as "*Uncharacterized*".)

***E. coli* genes**	**Other clades observed in**	**Orthologous group IDs**	**Evidence for regulation of gene expression**
yfhQ; suhB	5	COG0565; COG0483	SuhB has been implicated in the control of gene expression by modulating RNA turnover; the protein auto-regulates is own biosynthesis[24].
ffh; ypjD	5	COG0541; COG4137	YpjD is homologous to *Bacillus subtilis* HemX (full length hit), which negatively affects the steady-state cellular concentration of the HemA protein[25].
rplU; ispB	5	COG0261; COG0142	No evidence found yet; nevertheless RplU may (as a nucleic acid binding protein involved in translation[3]) be involved in expression regulation.
rrmJ; yhbY	5	COG0293; COG1534	No evidence found yet; however both proteins bind RNA, and three maize proteins harbouring domains similar to *yhbY* are required for chloroplast group II intron splicing[26, 27] – involvement in post-transcriptional regulation is thus possible.
gloB; yafS	5	COG0491; COG0500	YafS is a poorly characterized, putative methyltransferase[2, 3], which is in several species from other clades fused to transcriptional regulators of the ArsR family[2]. This may suggest at least indirect involvement in gene expression regulation.
rluD; yfiO	5	COG0564; COG4105	No evidence found yet. Nevertheless, as both proteins bind nucleic acids[2, 3], a regulatory role is conceivable.
yfjG; smpB	4	COG2867; COG0691	SmpB is essential for the activity of tmRNAs in *E. coli* and *Salmonella.* tmRNA initiates the bacterial salvage pathway of protein synthesis, and has been implicated in directly regulating gene expression[28].

yceF; yceD	4	COG0424; COG1399	No evidence found yet, however YceD is a poorly characterized, predicted nucleic acid binding protein[2, 3]; a regulatory role is thus conceivable.
gcp; rpsU	4	COG0533; COG0828	No evidence found yet; however RpsU could (as a nucleic acid binding protein involved in translation[3]) be involved in expression regulation.
ygcW; yqcE	4	COG1028; COG0477	no evidence found yet
rpH; yiCC	3	COG068; COG1561	RPH plays a role in tRNA metabolism, and moreover acts to regulate the attenuation of pyrE[29].
uvrA; ssB	3	COG0178; COG0629	SSB was recently shown to activate transcription of viral RNA polymerase promoters through template recycling[30]. It is thus possible that it also influences gene expression of bacterial genes.
map; rpsB	3	COG0024; COG0052	RpsB has been implicated to be indirectly involved in gene expression control: The protein is essential for binding of ribosomal protein S1 to the ribosome. S1 in turn is involved in functioning of tmRNAs, thus involved in the bacterial salvage pathway of protein synthesis[31], and possibly in regulation of gene expression[28].
radC; dfp	3	COG2003; COG0452	no evidence found yet
yraL; yraM	3	COG0313; COG3107	no evidence found yet, however both proteins are only very poorly characterized[3].
yjeK; efp	3	COG1509; COG0231	EFP stimulates efficient translation *in vitro,* is however not essential *in vitro*. It was thus suggested[32] that it might specifically regulate some mRNAs.
folB; ygiH	3	COG1539; COG0344	no evidence found yet
ybiA; ybjA	3	COG2963; COG2801	No evidence found yet; both proteins are predicted transposases residing on the *E. coli* F-Plasmid.

(Supplementary Table 5, continued)

8. Phylogenetic tree construction

A phylogeny of proteins assigned to KOG2969 was constructed using MRBAYES v2.01 (ref. 33), ignoring a short mouse protein fragment (accession Q9CT61). Manually refined alignments (pre-computed using CLUSTALW[34]) were used to derive tree topologies with maximum likelihood branch length estimates provided by TREE-PUZZLE[35]. MRBAYES was used with four heated chains over 250,000 generations. The likelihood of trees was examined to estimate the length of the burn-in phase, and all trees sampled 20,000 generations later than this point were used to create a consensus tree using 50% majority rule. Both MRBAYES and TREE-PUZZLE were used with the JTT model[36] of amino acid substitution, assuming the presence of invariant sites and using gamma distribution approximated by four different rate categories to model rate variation between sites, estimating amino acid frequencies from the alignment.

9. Data retrieval

DT-pairs with evolutionarily conserved gene organization, as well as orthologous groups[8] which have previously been classified as transcriptional regulators can be retrieved from the following website:
http://www.bork.embl.de/Docu/Bidirectional_genes/index.html

References of the Supplementary information

1. Pruess, M. et al. The Proteome Analysis database: a tool for the in silico analysis of whole proteomes. *Nucleic Acids Res.* **31**, 414-417 (2003).
2. von Mering, C. et al. STRING: a database of predicted functional associations between proteins. *Nucleic Acids Res.* **31**, 258-261 (2003).
3. Tatusov, R.L., Koonin, E.V. & Lipman, D.J. A genomic perspective on protein families. *Science* **278**, 631-637 (1997).
4. Kanehisa, M., Goto, S., Kawashima, S. & Nakaya, A. The KEGG databases at GenomeNet. *Nucleic Acids Res.* **30**, 42-46 (2002).
5. Salgado, H. et al. RegulonDB (version 3.2): transcriptional regulation and operon organization in Escherichia coli K-12. *Nucleic Acids Res.* **29**, 72-74 (2001).
6. Munch, R. et al. PRODORIC: prokaryotic database of gene regulation. *Nucleic Acids Res.* **31**, 266-269 (2003).
7. Madan Babu, M. & Teichmann, S.A. Evolution of transcription factors and the gene regulatory network in Escherichia coli. *Nucleic Acids Res.* **31**, 1234-1244 (2003).
8. Von Mering, C. et al. Genome evolution reveals biochemical networks and functional modules. *Proc. Natl. Acad. Sci. U S A* **100**, 15428-15433 (2003).
9. Tatusov, R.L. et al. The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* **4**, 41 (2003).
10. Salgado, H., Moreno-Hagelsieb, G., Smith, T.F. & Collado-Vides, J. Operons in Escherichia coli: genomic analyses and predictions. *Proc. Natl. Acad. Sci. U S A* **97**, 6652-6657 (2000).
11. Chandler, M.G. & Pritchard, R.H. The effect of gene concentration and relative gene dosage on gene output in Escherichia coli. *Mol. Gen. Genet.* **138**, 127-141 (1975).
12. Gollub, J. et al. The Stanford Microarray Database: data access and quality assessment tools. *Nucleic Acids Res.* **31**, 94-96 (2003).
13. Pedersen, A.G., Jensen, L.J., Brunak, S., Staerfeldt, H.H. & Ussery, D.W. A DNA structural atlas for Escherichia coli. *J. Mol. Biol.* **299**, 907-930 (2000).
14. Skovgaard, M. et al. The atlas visualisation of genome-wide information, Vol. 33. (Academic Press, London, UK; 2002).
15. Workman, C. et al. A new non-linear normalization method for reducing variability in DNA microarray experiments. *Genome Biol.* **3**, research0048 (2002).
16. Zhou, X., Kao, M.C. & Wong, W.H. Transitive functional annotation by shortest-path analysis of gene expression data. *Proc. Natl. Acad. Sci. U S A* **99**, 12783-12788 (2002).
17. Ghrist, A.C. & Stauffer, G.V. Promoter characterization and constitutive expression of the Escherichia coli gcvR gene. *J. Bacteriol.* **180**, 1803-1807 (1998).

18. Ibanez, E., Campos, E., Baldoma, L., Aguilar, J. & Badia, J. Regulation of expression of the yiaKLMNOPQRS operon for carbohydrate utilization in Escherichia coli: involvement of the main transcriptional factors. *J. Bacteriol.* **182**, 4617-4624 (2000).
19. Wu, W.F., Urbanowski, M.L. & Stauffer, G.V. Characterization of a second MetR-binding site in the metE metR regulatory region of Salmonella typhimurium. *J. Bacteriol.* **177**, 1834-1839 (1995).
20. Pradel, E. & Pages, J.M. The AcrAB-TolC efflux pump contributes to multidrug resistance in the nosocomial pathogen *Enterobacter aerogenes*. *Antimicrob. Agents. Chemother.* **46**, 2640-2643 (2002).
21. Heusipp, G., Schmidt, M.A. & Miller, V.L. Identification of rpoE and nadB as host responsive elements of Yersinia enterocolitica. *FEMS Microbiol. Lett.* **226**, 291-298 (2003).
22. Kovacikova, G. & Skorupski, K. The alternative sigma factor sigma(E) plays an important role in intestinal survival and virulence in Vibrio cholerae. *Infect. Immun.* **70**, 5355-5362 (2002).
23. Raina, S., Missiakas, D. & Georgopoulos, C. The rpoE gene encoding the sigma E (sigma 24) heat shock sigma factor of Escherichia coli. *EMBO J.* **14**, 1043-1055 (1995).
24. Inada, T. & Nakamura, Y. Autogenous control of the suhB gene expression of Escherichia coli. *Biochimie* **78**, 209-212 (1996).
25. Schroder, I., Johansson, P., Rutberg, L. & Hederstedt, L. The hemX gene of the Bacillus subtilis hemAXCDBL operon encodes a membrane protein, negatively affecting the steady-state cellular concentration of HemA (glutamyl-tRNA reductase). *Microbiology* **140 (Pt 4)**, 731-740 (1994).
26. Till, B., Schmitz-Linneweber, C., Williams-Carrier, R. & Barkan, A. CRS1 is a novel group II intron splicing factor that was derived from a domain of ancient origin. *RNA* **7**, 1227-1238 (2001).
27. Bateman, A. et al. The Pfam protein families database. *Nucleic Acids Res.* **30**, 276-280 (2002).
28. Withey, J.H. & Friedman, D.I. A salvage pathway for protein structures: tmRNA and trans-translation. *Annu. Rev. Microbiol.* **57**, 101-123 (2003).
29. Jensen, K.F. The Escherichia coli K-12 "wild types" W3110 and MG1655 have an rph frameshift mutation that leads to pyrimidine starvation due to low pyrE expression levels. *J. Bacteriol.* **175**, 3401-3407 (1993).
30. Davydova, E.K. & Rothman-Denes, L.B. Escherichia coli single-stranded DNA-binding protein mediates template recycling during transcription by bacteriophage N4 virion RNA polymerase. *Proc. Natl. Acad. Sci. U S A* **100**, 9250-9255 (2003).
31. Moll, I., Grill, S., Grundling, A. & Blasi, U. Effects of ribosomal proteins S1, S2 and the DeaD/CsdA DEAD-box helicase on translation of leaderless and canonical mRNAs in Escherichia coli. *Mol. Microbiol.* **44**, 1387-1396 (2002).
32. Aoki, H., Dekany, K., Adams, S.L. & Ganoza, M.C. The gene encoding the elongation factor P protein is essential for viability and is required for protein synthesis. *J. Biol. Chem.* **272**, 32254-32259 (1997).
33. Huelsenbeck, J.P. & Ronquist, F. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* **17**, 754-755 (2001).
34. Chenna, R. et al. Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Res.* **31**, 3497-3500 (2003).

35. Schmidt, H.A., Strimmer, K., Vingron, M. & von Haeseler, A. TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics* **18**, 502-504 (2002).

36. Jones, D.T., Taylor, W.R. & Thornton, J.M. The rapid generation of mutation data matrices from protein sequences. *Comput. Appl. Biosci.* **8**, 275-282 (1992).