THIS IS A PRELIMINARY VERSION – PLEASE DO NOT CITE OR REDISTRIBUTE

Title: Prediction and Validation of Operons in Uncharacterized Prokaryotes

Authors: Morgan N. Price, Katherine H. Huang, Eric J. Alm, and Adam P. Arkin

Author affiliation: Lawrence Berkeley Lab, Berkeley CA, USA

Corresponding author: Eric J. Alm, ejalm@lbl.gov

Date: February 26, 2004

**Abstract:** Operons have not been studied extensively outside of *Escherichia coli* and *Bacillus subtilis*. To predict operons in other prokaryotes, we combine comparative genomics predictions of conserved operons with probabilistic models of distances between genes in the same operon. Unlike previous efforts, which apply distance models from known *E. coli* operons to other organisms, we infer genome-specific distance models from the comparative genomics predictions and their estimated error rates. We validate our predictions against known operons from *E. coli* and *B. subtilis* and against microarray data for six diverse prokaryotes, testing whether adjacent genes predicted to be in the same operon (or not) are coexpressed. Genome-specific distance models for the archaeon *Halobacterium sp. NRC-1* and for *Helicobacter pylori* are significantly different from *E. coli*'s distance model, and we use microarray data to confirm these differences. Furthermore, *H. pylori* has many operons, contrary to earlier reports, and *Synechocystis sp. PCC 6803* has significant numbers of operons despite its unusual distance distribution. Finally, genomes with most of their genes on the leading strand of DNA replication have an even higher proportion of their multiple-gene transcripts on the leading strand. We use this observation to estimate the number of operons in strand-biased genomes and to improve our predictions significantly.

**Availability:** Operon predictions for 124 genomes, microarray similarity scores for six genomes, and source code in perl and R are available at http://vimssftp.lbl.gov/UnsupervisedOperons.

## Introduction

Operons are the fundamental unit of transcriptional regulation in prokaryotes, but little is known about operon structure outside of a few model organisms. Although over 100 complete prokaryotic genomes have been sequenced, efforts to predict operons computationally have focused on *Escherichia coli* and *Bacillus subtilis*. These methods rely on databases of experimentally identified transcripts for training and for validation (1, 2, 3, 4, 5, 6). As databases of known operons are not available for other species, unsupervised methods – methods that do not require large databases of known operons – are needed.

Two unsupervised methods have been proposed for predicting operons, based on (i) identifying operons that are conserved in multiple species or (ii) identifying pairs of genes on the same strand that have short non-coding regions between them. Genes that remain adjacent across long stretches of evolutionary time are almost certain to be in the same operon, but only half of *E. coli* operons can be identified this way (7). Genes that have short distances between them are usually in operons, and training data from known *E. coli* transcripts has been used to compute the relative likelihood of each intergenic distance for pairs that are within operons versus pairs that are at transcript boundaries. This probabilistic "distance model" is 82% accurate in *E.coli* (1). The *E. coli* distance model can also be used to predict operons in other organisms, but this *ad hoc* approach has been validated only in *B. subtilis* (8). As the distribution of intergenic distances within conserved operons varies across species (9), *E. coli* distance models will not be effective in many organisms.

To address the limited sensitivity of comparative approaches and the variation in distance distributions across species, we present an unsupervised learning approach to operon prediction that combines comparative genomics with a genome-specific distance model. We do not attempt to predict alternative transcripts due to internal promoters, terminator read-through, etc., as this remains a challenging problem even in *E. coli* (4), where transcriptional control features are relatively well characterized. Instead, we predict whether adjacent pairs of genes on the same strand can be transcribed together or not, and we use "transcription unit" and "operon" interchangeably.

Validation of operon predictions has been limited to *E. coli* and *B. subtilis*, again relying on databases of experimentally characterized transcripts. We test our unsupervised predictions against both microarray experiments and known operons in *E. coli* and *B. subtilis*, and against microarray data alone for four diverse prokaryotes where few operons have been characterized (*Helicobacter pylori*, *Chlamydia trachomatis*,

*Synechocystis sp. PCC 6803*, and *Halobacterium sp. NRC-1*).

# Methods

## Data sources

*Sequences.* We downloaded the complete annotated genomes of 124 prokaryotes from NCBI, TIGR, and DOE's JGI (Table S3), and excluded plasmids from our analyses.

*Known operons.* We obtained transcripts for *E. coli K12* from EcoCyc v7.5 (10) and for *B. subtilis* from http://cib.nig.ac.jp/dda/backup/taitoh/bsub.operon.html (11). We identified non-operon pairs as the same-strand pairs that are at the boundary of known transcripts, but are not within any known transcript (1). This gave 717 operon and 512 non-operon pairs for *E. coli* and 309 and 124, respectively, for *B. subtilis*.

*Microarrays.* We obtained data for *E. coli*, *B. subtilis*, and *H. pylori* from the Stanford Microarray Database (74, 78, and 31 arrays, respectively), for *Synechocystis* from the Kyoto Encyclopedia of Genes and Genomes (49 arrays), for *C. trachomatis* from T. Nicholson and R. Stephens (12 experiments times 2-3 replicates), and for *Halobacterium* from R. Bonneau and N. Baliga (44 arrays). See Table S4 for lists of individual experiments, including references, as well as analysis methods. To measure the similarity of the expression patterns of pairs of adjacent genes, we use the Pearson correlation ($r$) of normalized log-ratios.

## Overview of the unsupervised method

This section describes the key ideas behind the method (see Fig. 1 for an outline). To predict whether a pair of adjacent genes on the same strand is within an operon, the method examines (i) the distance between them in base pairs, (ii) several comparative genomics features, such as the extent to which this potential operon is conserved in other genomes, and (iii) the similarity of their codon adaptation index (CAI, (12)). The Features section (below) describes how we compute these features, why we used this set of features, and how we estimate likelihood ratios from features and training data. See Sup. Note 1 for mathematical details.

The unsupervised method relies on a key assumption, which we justify in the Results: the distribution of values of the comparative genomics features is the same for non-operon pairs (on the same strand) and for opposing-strand pairs. The method also requires an estimate of the total number of operons or operon

pairs in the genome, which is described in the next section.

The first step of the method is to make preliminary predictions using the comparative genomics features. We use adjacent pairs on opposing strands of DNA as a "true negative" set. By assumption, these have the same distributions as non-operon pairs. For each genome, we train a classifier to distinguish this "true negative" set from the adjacent same-strand pairs, which are a mixture of operon and non-operon pairs. This classifier converts the comparative genomics features into likelihood ratios of the pair being from the "mixture" set versus the "true negative" set. From the estimated number of operons, we know the proportion of true positives in the mixture set, so we can convert these likelihood ratios into the likelihood of each same-strand pair being in an operon or not. We also use a threshold on this likelihood to make preliminary predictions of whether same-strand pairs are within operons.

We train a genome-specific distance model from these preliminary predictions together with an estimate of their error rates, which can be derived from our assumption. The false positive rate – the error rate on same-strand non-operon pairs – equals the error rate on opposing-strand pairs (that is, the fraction of opposing-strand pairs that would be predicted to be in operons). The false negative rate is implied by the shortage of predicted operon pairs relative to expectations, which equals the estimated number of operon pairs plus the number of false positives minus the number of predicted operon pairs.

We combine likelihood ratios of same-strand pairs being in an operon (or not) from comparative genomics, from the genome-specific distance model, and from the proportion of same-strand pairs that are in operons, using Bayes' rule. This produces a second set of intermediate predictions, which we use to train a classifier for our final feature, the similarity of CAI, giving a fourth set of likelihood ratios. We do not attempt to estimate error rates at this step. Finally, we combine all four likelihood ratios, again using Bayes' rule, to give the probability of a pair being in an operon based on all the features. Our predicted operon pairs are those that are more likely to be in an operon than not ($p > 0.5$).

For *E. coli* and *B. subtilis*, we compare these unsupervised predictions to supervised predictions based on the same features, training the same classifier with experimentally known operon and non-operon pairs. We measure the accuracy of the supervised predictions using 100-fold cross-validation.

## The number of operons in strand-biased genomes

To estimate the number of operons in each genome, we extend "directon counting" to genomes with coding strand bias. In genomes without coding strand bias, where genes are equally likely to be on the leading

and lagging strands of DNA replication, non-operon pairs should be equally likely to be on the same strand or not. Then there are twice as many operons (including single-gene transcripts) as runs of same-strand pairs, or "directons" (7, 13).

Some genomes, including *B. subtilis*, have significantly more genes on the leading stand of replication. This is probably to avoid collisions between RNA and DNA polymerases that would result if the polymerases moved in opposite directions (14). In strand-biased genomes, we cannot simply use directon-counting, as both members of a non-operon pair are likely to be on the same strand (the leading strand).

If we know the relative frequency of multi-gene operons on the two strands, however, then we can correct for the strand bias. A simple model of strands and operons in strand-biased genomes has three parameters: the proportion of leading-strand pairs that are in operons, the proportion of lagging-strand pairs that are in operons, and the proportion of genes that are on the leading strand. We observe two variables: the proportion of pairs that are on the same strand, and the proportion of genes on the leading strand (estimated below). With three unknowns and two observables, we need one additional constraint to solve for the proportion of same-strand pairs that are in operons. We assume that same-strand pairs on either strand are equally likely to be operons ("strand-wise" estimate), or alternatively, that transcripts assort between strands independently of their length ("strand-naive" estimate). We use the "strand-wise" assumption for predictions (justified in the Results). For formulas, see Sup. Note 1.

We estimate strand bias in each genome by finding the minimum and maximum of the cumulative strand bias of genes along the chromosome. These extrema usually approximate the origin and terminus of replication, but other factors can also create coding strand bias. We estimate 74% of genes on the leading strand for *B. subtilis* and 56% for *E. coli*, close to the values of 74% and 54% given by (14).

## Features

For each pair of adjacent genes, we compute:

Several "gene neighbor" scores measuring how conserved the adjacent pair is

Whether the two genes have the same COG function class (15)

The distance between them (the length of the intergenic region of DNA)

The similarity of the codon adaptation index (CAI, (12)) for the two genes

For formulas and implementation details, see Sup. Note 1.

The gene neighbor method measures how often two genes are near each other across many genomes (16, 17). We use putative orthologs from bidirectional best hits, and ask how often the genes have orthologs that are within 5 kb. Previous workers threw out closely related genomes (17) or reduced sensitivity when they were present (7). Instead, we cluster related genomes together (see Table S3) and compute several scores over different sets of clusters. To get useful information from closely related genomes, we introduce a penalty if both orthologs exist but are not within 5 kb.

We considered using phylogenetic profiles (18) as well, but found no benefit (see Results). The similarity of textual annotations has been used to select a genome-specific distance threshold, but this threshold and the underlying feature were used to aid functional annotation, and their effectiveness for operon prediction was not directly tested (19). This feature and other precise measures of functional similarity (2) might further improve the comparative predictions.

We use a codon adaptation-based feature (CAI) because genes in the same operon are often expressed at similar levels. Similarity of codon usage is an informative feature for predicting operons in *E. coli* (5).

Given training sets with errors, we use each feature to classify adjacent pairs. Specifically, we estimate log likelihood ratios from (i) a training set split into two classes, such as conserved and non-conserved pairs, (ii) values of the feature for each pair in the training set, and (iii) error rates. We first transform the feature into ranks. The likelihood ratio for a range of ranks (e.g., a bin in a histogram) is the ratio of counts for the two classes within that range, corrected for the error rates and the unequal representation of the two classes in the training set. To avoid overfitting, we use a generalization of pseudocounts within each range, and smooth the log likelihood ratios across ranges by using local regression (Sup. Note 1).

The comparative genomics predictions use the gene neighbor scores and the similarity of COG function class. The other features are sensitive to strandedness, and cannot be used until a later step of the method where only same-strand pairs are being considered. As the comparative features are highly correlated, we find the best-fitting linear combination of log likelihood ratios with logistic regression (*glm* in R, http://www.r-project.org/) instead of using Bayes' rule.

# Results

## Surprisingly few operons on the lagging strand

The unsupervised method requires a prior estimate of the number of operons in the genome. To estimate this quantity for genomes with coding strand bias, we need to know the relative proportion of multi-gene and single-gene transcripts on the leading and lagging strands. A reasonable null hypothesis is that transcription units assort between the leading and lagging strands independently of their length. In this case, non-operon pairs on the lagging strand should be rare: given a first transcript on the lagging strand, the probability of the next transcript being on the lagging strand independently is low. Hence, same-strand pairs on the lagging strand should be more likely to be in operons than pairs on the leading strand (for a rigorous proof see Sup. Note 1). In *B. subtilis*, however, the distributions of intergenic distances for leading and lagging (same-strand) pairs are remarkably similar (Kolmogorov-Smirnov D-statistic $= 0.03$, $p = 0.71$). Furthermore, the proportion of pairs that are conserved within 5 kb in a distant genome is much higher on the leading strand (33.3% vs. 16.1%, Fisher exact test $p < 10^{-14}$). In general, strand-biased genomes show no consistent difference between the distributions of distances for the two strands, and both strand-biased and non-biased genomes show a strong preference for conserved gene neighbors on the leading strand (Fig. 2).

Our interpretation is that same-strand pairs on the lagging strand are about as likely to be in operons as same-strand pairs on the leading strand, leading to the similarity in the distance distributions, whereas highly conserved operons are selected to the leading strand in all genomes. This "strand-wise" hypothesis gives better operon predictions than the "strand-naive" null hypothesis: (i) more accurate predictions in *B. subtilis*, (ii) better distance models, and (iii) better agreement with *E. coli*-based estimates of the number of operons. First, in *B. subtilis*, the strand-wise approach gives a higher estimate of the proportion of same-strand pairs that are within operons – 0.517 vs. 0.413 – that leads to significantly better agreement with both known operons and microarrays. The area under the operating curve (Fig. 3, middle panel) is 0.888 for strand-wise and 0.864 for strand-naive ($p = 2 \cdot 10^{-6}$, DeLong test (20)). This corresponds to accuracy at the default threshold of 81.0% and 78.0%, respectively (computed from the mean of the accuracy on known operon and non-operon pairs). The agreement with microarrays (Spearman correlation of predicted $p$ with microarray similarity $r$) is 0.461 and 0.433, respectively ($p < 10^{-10}$ from two-sided $t$-test of correlation between $rank(r)$ and the differences in $rank(p)$). Second, across 124 genomes the strand-

wise estimates leads to distance models more in accord with expectations from *E. coli* and *B. subtilis* (Fig. S6). Finally, strand-wise estimates of the number of operons agree better with estimates from *E. coli* distance distributions (the method of (8)). The Spearman correlation with *E. coli*-based estimates is 0.363 for strand-wise and 0.223 for strand-naive estimates ($p = 0.04$ from correlation test of ranked differences). This poor agreement between our method and the *E. coli*-based method (see Fig. S8B) may reflect biologically meaningful variation in the distance distributions of different genomes (9).

## Non-operon pairs resemble opposing-strand pairs

The unsupervised method assumes that comparative genomics features will have the same distributions for non-operon pairs as for opposing-strand pairs. We test the equivalent assertion that the distribution of $p$-values from the comparative predictions will be the same, using putative non-operon pairs at the boundaries of known operons in *E. coli* or *B. subtilis* (see Methods). In *E. coli*, the distributions are quite similar (Kolmogorov-Smirnov D-statistic = 0.12, $p = 1.1 \cdot 10^{-4}$), and many conserved putative non-operon pairs are actually known to be co-transcribed (7). *B. subtilis*, however, shows a much larger difference (D-statistic = 0.31, $p = 8.0 \cdot 10^{-10}$), because of conserved non-operon pairs (Fig. S7). We checked the 19 non-operon pairs that strongly disagreed with comparative predictions (those with predicted $p > 0.9$) against transcription unit diagrams and Northern hybridizations at BSORF (http://bacillus.genome.ad.jp/bsorf.html). Northerns were only available for three pairs (sul/folA, mmgE/yqiQ, and deoR/dra), and in all three cases, there was a transcript containing both genes. Furthermore, in both *E. coli* and *B. subtilis*, the cases where unsupervised predictions disagree with known operons (in either direction) show similar and intermediate levels of coexpression relative to the operon and non-operon pairs where both agree (Fig. 3). Thus, apparent deviations from the assumption reflect the limitations of the set of known non-operon pairs, perhaps due to alternative transcripts.

## Comparison to supervised methods

In *E. coli* and *B. subtilis*, the unsupervised predictions are about as accurate as the supervised predictions, and are competitive with previously published (supervised) predictions. The unsupervised method has sensitivity and specificity at the default threshold ($p > 0.5$) of 88.3% and 79.9% in *E. coli* and 90.9% and 71.0% in *B. subtilis* (for comparisons to published methods see Table S1). In *E. coli*, the area under the operating curve (middle panel of Fig. 3) is 0.920, versus 0.919 for the supervised method ($p = 0.34$,

DeLong test). In *B.subtilis*, the areas are 0.888 and 0.907, respectively ($p = 0.02$, DeLong test). In *E.coli*, the agreement with microarrays (the Spearman correlation of predicted $p$ with microarray similarity $r$) is 0.494, versus 0.499 for the supervised method ($p = 0.19$ from ranked difference correlations). In *B. subtilis*, agreement is 0.461 and 0.489, respectively ($p = 0.001$). The unsupervised distance models are similar to the supervised models (Fig. 3), and on known operons, unsupervised $p$-values are consistent with prediction accuracy (data not shown).

## Accuracy against microarray data

To test operon predictions more broadly, we compare the unsupervised predictions to microarray data from six species. We test whether adjacent genes predicted to be in the same operon (or not) have similar expression patterns as measured by the Pearson correlation ($r$). We also use the microarray data to estimate the accuracy of those predictions and to test the genome-specific distance models.

In all six species, the unsupervised predictions correlate with the similarity of expression (Fig. 4). The agreement of predictions with individual microarrays is roughly consistent across species (Fig. S10). The method combines features effectively: combining comparative genomics with intergenic distance improves accuracy over either measure alone, and the combined comparative predictions outperform the best single comparative feature in five of the six species (Table S2). Similarity of codon adaptation has little impact on the results (Table S2).

To estimate prediction accuracy, we model the distribution of microarray similarity for predicted operon and non-operon pairs as mixtures of the true operon and non-operon distributions. To estimate the distribution for true operon pairs, we assume that high-confidence predictions (pairs with predicted $p > 0.95$) are reliable. Unsupervised $p$-values are in agreement with the accuracy of predictions for known operons in *E. coli* and *B. subtilis* (data not shown), and in all six species the average microarray similarity ($r$) rises sharply as $p$ approaches 1 (Fig. 4). To estimate the distribution for non-operon pairs, we once again assume that non-operon pairs will look like opposing-strand pairs. We then estimate the proportion of true operons within our predicted sets by fitting a mixture of densities, and rerun the estimates on subsets of experimental conditions to get confidence intervals (see Fig. S9 for details). We compare these estimates to the accuracy expected from the predicted $p$-values. The overall accuracy from microarrays (the average of the false positive and false negative rates) is consistent with expectations in *E. coli* and *C. trachomatis*, and slightly lower than expected in *B. subtilis* (79% $\pm$ 1% vs. 84%), while there is insufficient data for

reliable estimates in *H. pylori* or *Halobacterium* (Fig. S9). Although overall accuracy in *Synechocystis* is consistent with expectations (72% ± 5% vs. 73%), this reflects the combination of a high false positive rate and a low false negative rate.

*Synechocystis has fewer operons than predicted.* The microarray-based estimate of the proportion of same-strand pairs that are in operons is $0.34 \pm 0.06$, compared to 0.48 from strand-wise or 0.49 from "directon counting" (this genome has little coding strand bias). *Synechocystis* is known to have unusual operons, with a very wide distribution of distances between conserved pairs (see (8, 9) and Fig. 5). It has been proposed that errors in gene models, specifically the absence of TTG initiation codons in the predicted gene starts, might create this discrepancy (8). This would not affect our estimate of the number of operons, however, as it does not depend on intergenic distances. Furthermore, using other gene start predictions does not change the distance distributions (data not shown; alternative gene models were computed with CRITICA (21), which led to 5% TTG starts, or downloaded from CyanoBase, http://www.kazusa.or.jp /cyano/). Pseudogenes or unannotated ORFs could also create large distances within conserved operons, but *Synechocystis* contains very few intergenic regions with homology to annotated ORFs (Sup. Note 2). Thus, both the wide distribution of distances between conserved pairs and the apparent surplus of same-strand pairs that are not in operons remain a mystery. Nevertheless, both the unsupervised method and the genome-specific distance model are effective for this organism (Fig. 4 and Table S2).

*Microarrays confirm that distance models vary.* In *Halobacterium*, pairs separated by relatively short distances (25-50 bp) are unlikely to be conserved in a distant genome (12/190 = 6.3% vs. 173/1021 = 16.9% for other pairs, $p < 10^{-4}$, Fisher exact test), whereas in *E. coli* such pairs are as likely to be conserved as other pairs (100/254 = 39.4% vs. 983/2751 = 35.7%, $p$=0.25). Thus, the genome-specific distance model deviates significantly from the supervised *E. coli* model in its predictions. The *Halobacterium* distance model shows better agreement with microarrays: the Spearman correlation of the difference between binary predictions (using the default threshold of $p > 0.5$) with $r$ is 0.08 ($p = 0.008$; for the other five genomes $p > 0.05$). Similarly, in *H. pylori*, pairs separated by 50-100 bp are significantly less likely to be conserved than in *E. coli*: 95% confidence intervals for odds ratios are $0.15 - 0.49$ in *H. pylori* and $0.50 - 0.80$ in *E. coli*. Thus, pairs at this distance have lower probabilities of being operons according to the genomic-specific distance model (Fig. 5). Microarray data suggests that these 143 pairs contain few operons, while the corresponding pairs from *E. coli* do contain operons. *H. pylori* pairs separated by 50-100 bp have the same distribution of $r$ as do other predicted non-operons (Kolmogorov-Smirnov D-

statistic = 0.06, $p > 0.5$), and a significantly different distribution from other predicted operons ($D = 0.33$, $p < 10^{-11}$). The corresponding tests in *E. coli* give $D = 0.24$ and 0.23, respectively.

*Large numbers of operons in the $\epsilon$-Proteobacteria.* It has been suggested that *H. pylori* and its relative *Campylobacter jejuni* have few operons (22, 23). Our strand-wise method predicts that most same-strand pairs in these genomes are in operons – 71% in *H. pylori* and 72% in *C. jejuni*, higher than for *E. coli* or *B. subtilis*. Conserved same-strand pairs are separated by smaller distances in both genomes: the Spearman correlations of $p$-values from comparative genomics with intergenic distance are $-0.27$ and $-0.17$, respectively. The two species share large numbers of adjacent pairs, which are probably conserved ancestral operons: 20.5% of same-strand adjacent pairs in *H. pylori* are conserved within 5 kb in *C.jejuni*, versus only 3.4% of opposing-strand pairs ($p < 10^{-13}$, $\chi^2$ test). Finally, and most significantly, in microarray data for *H. pylori*, predicted operon pairs have much greater similarity than predicted non-operon pairs (Fig. 4).

## Predicted operons across 124 genomes

To test the predictions for 124 genomes, where microarray data is generally not available, we examine the the internal consistency of the unsupervised method, the genome-specific distance models, and an internal estimate of accuracy. The first step of the unsupervised method is to train comparative genomics features to distinguish same-strand from opposing-strand pairs. The method achieves significant discrimination between the two sets in all 124 genomes (Fig. S8A), and the extent of discrimination is strongly correlated with the surplus of conserved same- vs. opposing-strand pairs (Spearman correlation = 0.59, $p = 4.9 \cdot 10^{-13}$). Combining multiple gene neighbor scores and adding COG functional classes improves the discrimination significantly, but using phylogenetic profiles of gene co-occurrence in genomes does not (generalized analysis of variance, Fig. S8D). Within same-strand pairs, the agreement of the comparative predictions with distance-only predictions is greater than the agreement between raw gene neighbor scores and distances for 112/124 genomes (Fig. S8C).

Most genome-specific distance models have the shape expected from *E. coli* and *B. subtilis*, but *E. coli* has particularly extreme values at very short and very high separations (Fig. 5). *E. coli* may have an unusually strong correlation between intergenic distance and conserved proximity, or gene starts in other genomes may simply be less accurate (e.g., (8)).

*Pseudogenes in ancestral operons.* The correlation between intergenic distance and conserved proximity

might be weakened in some genomes by the disruption of genes within ancestral operons. For example, *B. anthracis str. Ames* has an unusual distance model, while its relative *B. subtilis* has a typical model (Fig. 5). By examining pairs of adjacent genes whose intergenic regions have syntenic homology to annotated genes in *B. cereus* (a close relative of *B. anthracis*), we estimate that 7% of same-strand pairs in *B. anthracis* contain potential pseudogenes, remnants of truncated genes, or erroneous start predictions (Sup. Note 2). *B. anthracis* also has 12 apparent pseudogenes within highly conserved operons, compared to none in *B. subtilis* and two in *E. coli*. We examined two of these pseudogenes in *B. anthracis str. A2012*, and found that the reading frames was also disrupted, so these are unlikely to be sequencing errors (Sup. Note 2). As operons containing pseudogenes would likely be disrupted by polarity (rho-dependent termination or mRNA decay), pseudogenes could also affect the estimated number of operons.

*Accuracy in 124 genomes.* The estimated accuracy, from the average difference between the $p$-values and zero or one, ranges from 71% to 96%, with half of the values lying between 82% and 87% (Fig. S11). As accuracy is strongly correlated with the surplus of conserved same- vs. opposing-strand pairs (Spearman correlation 0.47, p=$3 \cdot 10^{-8}$), accuracy may improve as more genomes are sequenced.

## Conclusions

Interpreting the wealth of microbial sequence data requires unsupervised methods for statistical inference and careful validation against experiment across diverse species. We demonstrate accurate unsupervised prediction of operons by combining comparative genomics and genome-specific distance models. The method is based on an assumption, which we validate against known operons, that the evolution of same-strand pairs which are not in operons resembles that of opposing-strand pairs.

We use microarray data to show that unsupervised predictions are effective in phylogenetically diverse prokaryotes, including the archaeon *Halobacterium NRC-1*, a Gram-positive (*B. subtilis*) with strong coding strand bias, a member (*H. pylori*) of the $\epsilon$-proteobacteria, which have been described as having few operons (22, 23), and *Synechocystis PCC 683*, which has unusual operons (8, 9). Furthermore, in *E. coli* and *B. subtilis*, unsupervised predictions are about as accurate as supervised predictions.

It has been proposed that distributions of distances within operons across the prokaryotes are similar to *E. coli*, and that this can be used to predict individual operons and to estimate the total number of operons (8). Distance models from *E. coli* disagree significantly with patterns of conservation in *H. pylori*

and *Halobacterium*, and microarray data for those two organisms confirms these differences. Estimates of the number of operons from *E. coli* distance models do not agree with estimates from counting same-strand pairs or "directons."

As *B. subtilis* and other genomes with coding strand bias have similar distributions of distances between same-strand pairs on the leading and lagging strands, we infer that same-strand pairs on both strands are equally likely to be in operons. We use this observation to generalize "directon counting" to to genomes with coding strand bias, and to improve our operon predictions. Coding strand bias is believed to reflect avoidance of collisions between DNA and RNA polymerases (14). We speculate that there is a balance between greater selection for multi-gene transcripts on the leading strand and a lower probability of adjacent non-operon pairs forming on the lagging strand by chance. This selection could reflect more frequent collisions between RNA and DNA polymerases on longer transcripts or higher expression levels of multi-gene transcripts.

Our present method relies largely on conserved gene neighbors and on intergenic distance. We improve the sensitivity of the gene neighbor method by handling distantly and closely related species separately and by introducing a penalty if both orthologs are present but are not near each other. Phylogenetic profiles do not provide statistically significant additional information after combining several gene neighbor scores and considering whether COG functional classes match. One approach to identifying further features would be to investigate the contents of the conserved but widely separated adjacent pairs found in some genomes, such as the apparent pseudogenes in *B. anthracis* and the continuing mystery in *Synechocystis*. We also suspect that comparative identification of the features that determine whether genes are co-transcribed, such as conserved transcription initiation sites or rho-independent terminators, could be effective.

## Acknowledgments

We thank Nitin Baliga, Richard Bonneau, Tracy Nicholson, and Richard Stephens for providing unpublished microarray data, Inna Dubchak for insightful discussions and sharing resources, and the Arkin lab for comments on the manuscript. This work was supported by a grant from the DOE Genomes To Life program (DE-AC03-76SF00098).

## References


1. Salgado, H., Moreno-Hagelsieb, G., Smith, T.F., & Collado-Vides, J. (2000) *Proc. Natl. Acad. Sci. USA* **97**:6652–7.

2. Zheng, Y., Szustakowski, J.D., Fortnow, L., Roberts, R.J., & Kasif, S. (2002) *Genome Res.* **12**:1221–30.

3. Sabatti, C., Rohlin, L., Oh, M.K., & Liao, J.C. (2002) *Nucleic Acids Res.* **30**:2886–93.

4. Bockhorst, J., Qiu, Y., Glasner, J., Liu, M., Blattner, F., & Craven, M. (2003a) *Bioinformatics* **19 Suppl. 1**:I34–I43.

5. Bockhorst, J., Qiu, Y., Glasner, J., Liu, M., Blattner, F., & Craven, M. (2003b) *Bioinformatics* **19**:1227–35.

6. de Hoon, M.J.L., Imoto, S., Kobayashi, K., Ogasawara, N., & Miyano, S. (2004) *Pac. Symp. Biocomputing* 2004.

7. Ermolaeva, M.D., White, O., & Salzberg, S.L. (2001) *Nucleic Acids Res.* **29**:1216–21.

8. Moreno-Hagelsieb, G. & Collado-Vides, J. (2002) *Bioinformatics* **18 Suppl. 1**:S329–36.

9. Rogozin, I.B., Makarova, K.S., Natale, D.A., Spiridonov, A.N., Tatusov, R.L., Wolf, Y.I., Yin, J., & Koonin, E.V. (2002) *Nucleic Acids Res.* **30**:4264–71.

10. Karp, P.D., Riley, M., Saier, M., Paulsen, I.T., Collado-Vides, J., Paley, S.M., Pellegrini-Toole, A., Bonavides, C., & Gama-Castro, S. (2002) *Nucleic Acids Res.* **30**:56-8.

11. Itoh, T., Takemoto, K., Mori, H., & Gojobori, T. (1999) *Mol. Biol. Evol.* **16**:332–46.

12. Sharp, P.M. & Li, W.-H. (1987) *Nucl. Acids Res.* **15**:1281–1295.

13. Cherry, J.L. (2003) *J. Theor. Biol.* **221**:401–10.

14. Rocha, E.P.C. (2002) *Trends Microbiol.* **10**:393–5.

15. Tatusov, R.L., Natale, D.A., Garkavtsev, I.V., Tatusova, T.A., Shankavaram, U.T., Rao, B.S., Kiryutin, B., Galperin, M.Y., Fedorova, N.D., & Koonin, E.V. (2001) *Nucleic Acids Res.* **29**:22–8.

16. Overbeek, R., Fonstein, M., D'Souza, M., Pusch, G.D., & Maltsev, N. (1999) *Proc. Natl. Acad. Sci. USA* **96**:2896–901.

17. Huynen M., Snel, B., Lathe 3rd, W., & Bork, P. (2000) *Genome Res.* **10**:1204–10.

18. Pellegrini, M., Marcotte, E.M., Thompson, M.J., Eisenberg, D., & Yeates, T.O. (1999) *Proc. Natl. Acad. Sci. USA* **96**:4285-8.

19. Strong, M., Mallick, P., Pellegrini, M, Thompson, M.J., & Eisenberg, D. (2003) *Genome Biol.* **4**:R59.

20. DeLong, E.R, DeLong, D.M., & Clarke-Pearson, D.L. (1988) *Biometrics* **44**:837–45.

21. Badger, J. H. & Olsen, G. J. (1999) *Mol. Biol. Evol.* **16**: 512-524.

22. Thompson, L.J., Merrell, D.S., Neilan, B.A., Mitchell, H., Lee, A., & Falkow, S. (2003) *Infect Immun.* **71**:2643–55.

23. Parkhill, J., Wren B,W., Mungall, K., Ketley, J.M., Churcher, C., Basham, D., Chillingworth, T., Davies, R.M., Feltwell, T., Holroyd, S. & others. (2000) *Nature* **403**:665–8.

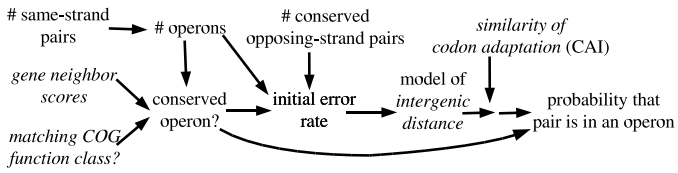



Figure 1: Overview of our unsupervised method to predict operons. Features are shown in italics.

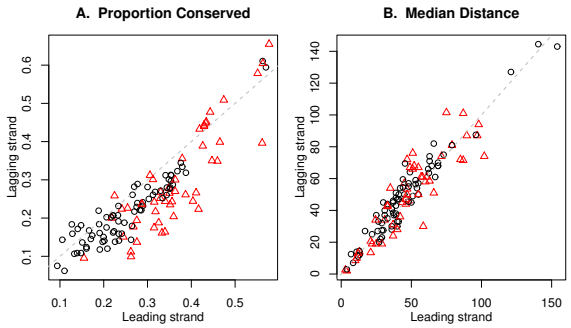



Figure 2: Comparison of leading and lagging strands across 124 genomes. (A) Proportion of same-strand pairs conserved in a distant genome. (B) Median distance between same-strand pairs. The leading strand is on the $x$ axes, the lagging strand is on the $y$ axes, the dashed grey lines show $x = y$, and red triangles indicate genomes with 60% or more of their genes on the leading strand.

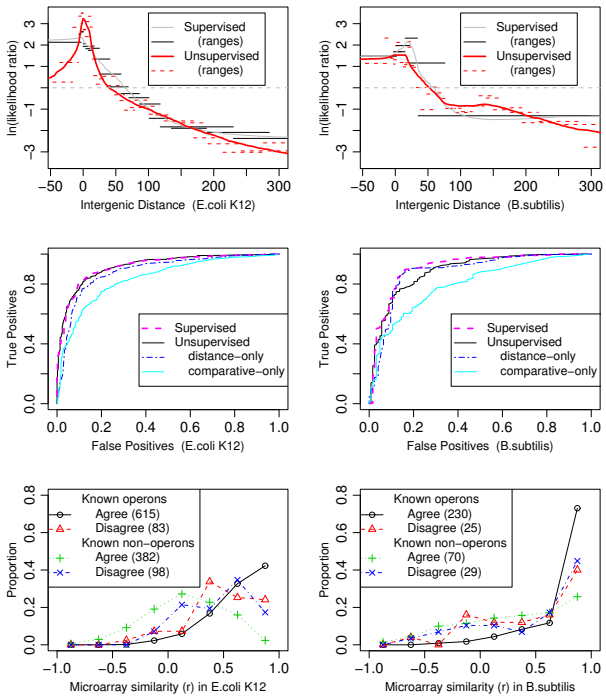



Figure 3: Effectiveness of unsupervised predictions in *E. coli K12* (left) and *B. subtilis* (right). Top row: Distance models (log likelihood ratios), both supervised and unsupervised, showing the smoothed models used in our predictions (curves) and the raw log likelihood ratio for each range (horizontal lines). If the log likelihood ratio is zero (indicated by the dashed horizontal line), then that value of distance is equally likely for operon and non-operon pairs. Middle row: Accuracy on known operon and non-operon pairs as the prediction threshold varies, also known as the receiver operating characteristic curve. The area under the curve is the probability that a known operon pair will have a higher score than a known non-operon pair if both pairs are chosen randomly. Bottom row: Distribution of microarray similarity (Pearson correlation $r$) for pairs where unsupervised predictions agree or disagree with known operons.

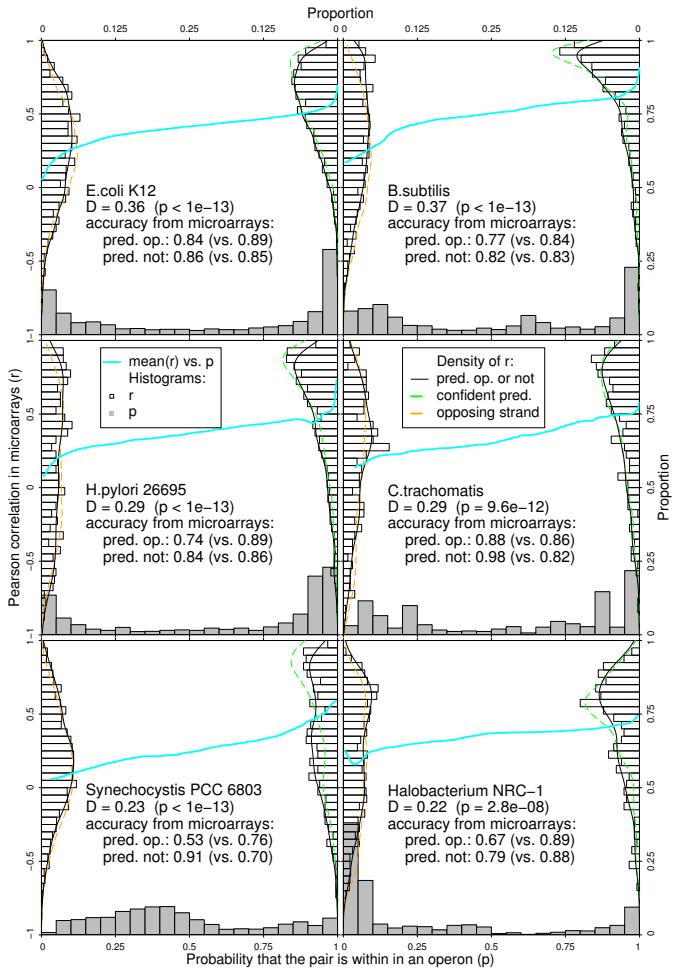



Figure 4: Agreement of predicted operons with microarray data for six species. Each panel shows the distribution of microarray similarity (Pearson $r$) for predicted operon pairs (right of panel) and non-operon pairs (left of panel). The cyan curve in the middle of each panel shows the smoothed average of $r$ as a function of the predicted probability of being in an operon ($p$) for all same-strand pairs. The histogram at the bottom shows the distribution of $p$. We also report the Kolmogorov-Smirnov test of whether the two distributions of $r$ differ, and the accuracy of both predicted sets, as estimated from microarrays or implied by the $p$-values (shown in parentheses). The accuracy estimates compare the kernel densities of $r$ for the predicted sets (black curves) to those for opposing-strand pairs (orange dashed curves, left) and high-confidence predicted operon pairs (green dashed curves, right). The smoothed average of $r$ vs. $p$ is estimated by local regression of $r$ vs. $rank(p)$.

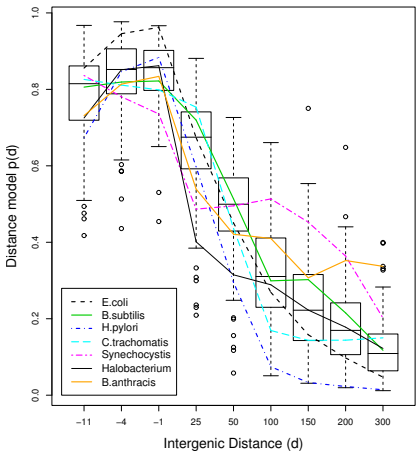
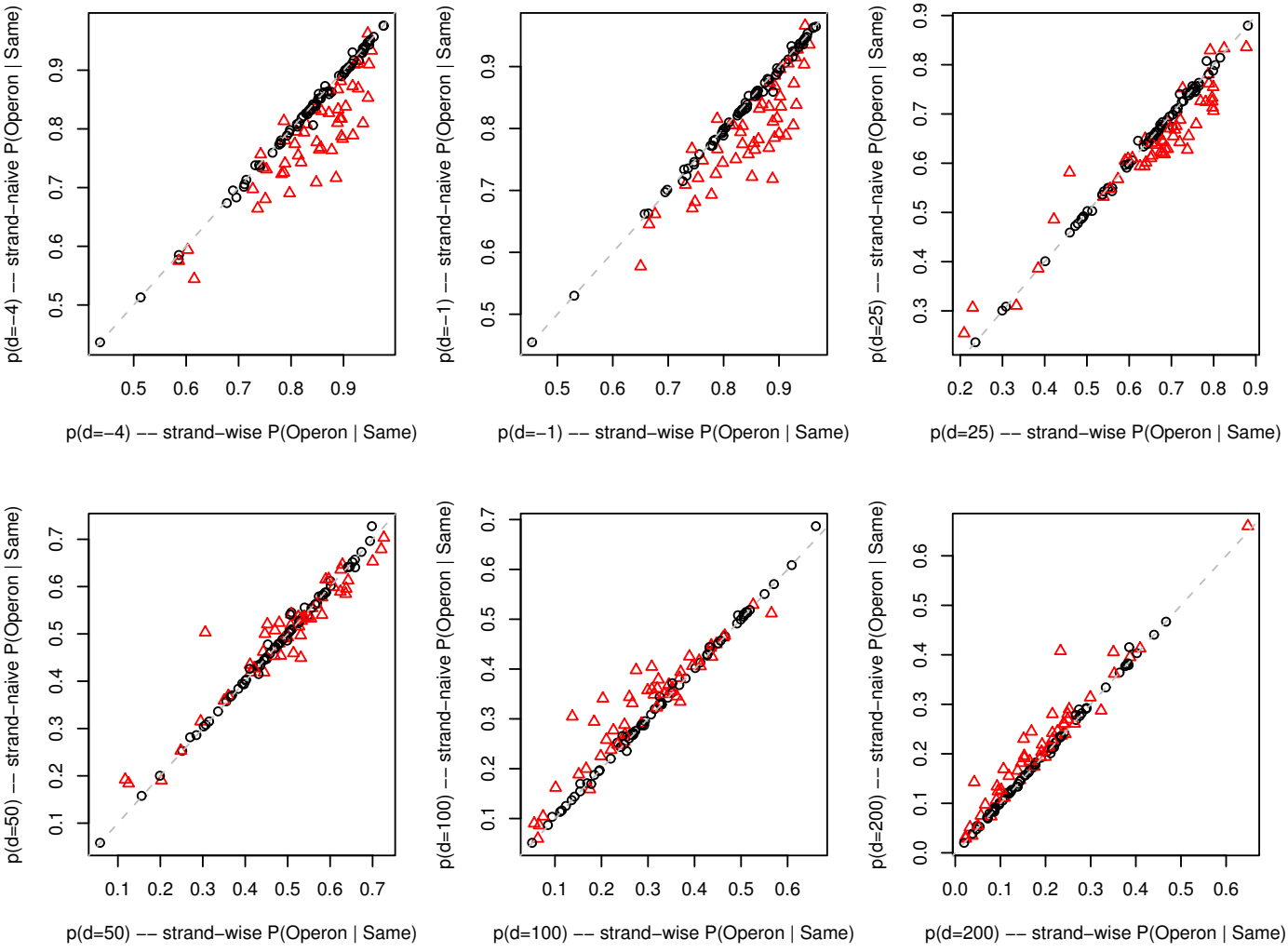
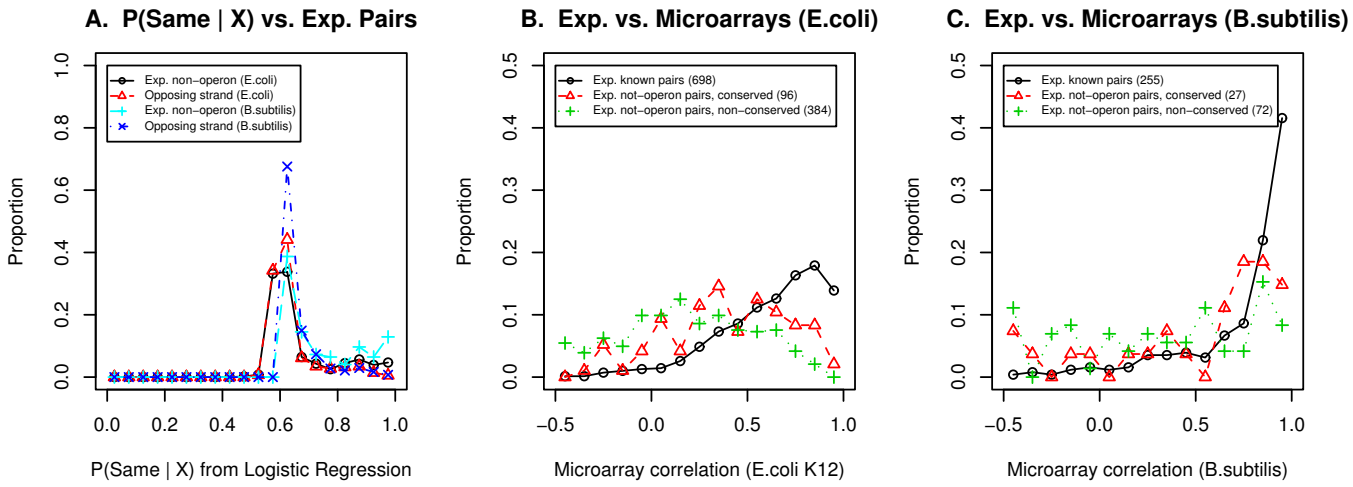


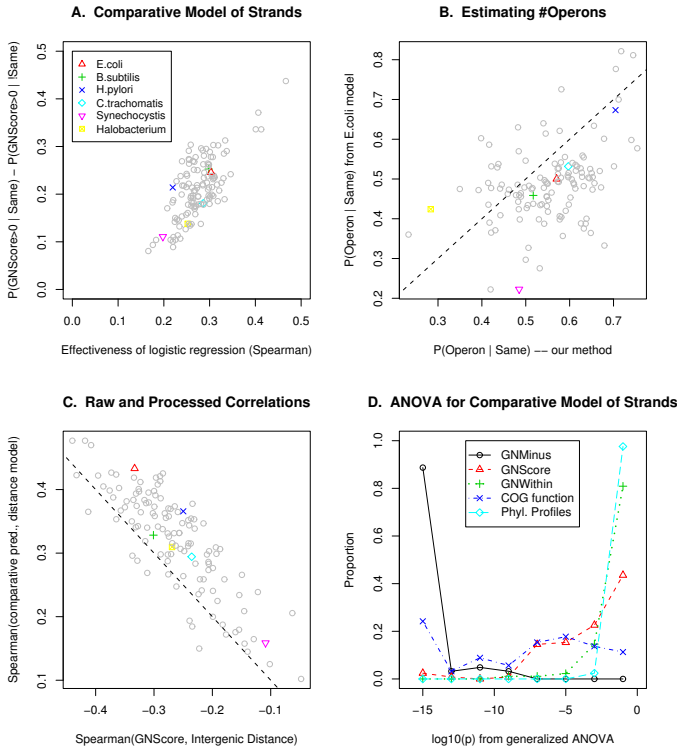

Figure 5: Unsupervised distance models ($p_d$) across 124 genomes. (If $p_d = .5$, operon- and non-operon pairs are equally likely to have that value of distance.) The boxes show quartiles and medians, whiskers extend up to 1.5x the interquartile range from the box, and dots show outliers. Note the non-linear $x$ axis.




**Supplementary Figure 6: Effect of the proportion of same-strand pairs that are in operons, $P(Operon|Same)$, on the distance model $p_d$.** The $x$-axis shows results when assuming that $P(Operon|Same)$ is the same on both strands (the "strand-wise" method used for predictions), and the $y$-axis shows results when assuming that transcriptional units are assorted between strands independently of the whether or not they contain more than one gene ("strand-naive"). Genomes with strong strand bias (>60% of the genes on the leading strand) are shown with red triangles, and the dashed grey lines show $x = y$.



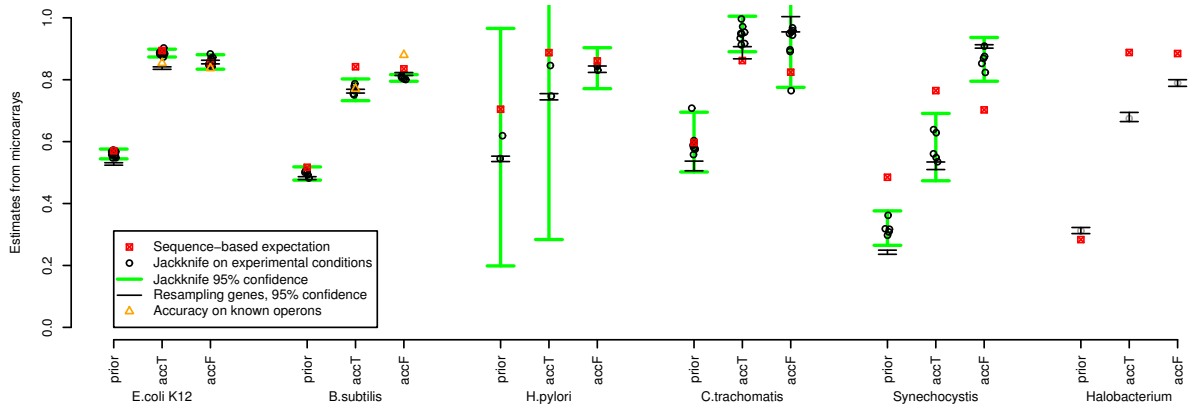

**Supplementary Figure 7: Direct tests of our assumption.** (A) Distribution of $p$-values from our comparative model of same-strand vs. opposing-strand pairs, $P(Same|\vec{X})$, for putative non-operon pairs from experiments and for opposing-strand pairs, in *E. coli K12* and in *B. subtilis*. (B and C) Distribution of microarray correlations for experimentally identified same-operon pairs, for known non-operon pairs conserved within 5 kb in a distant genome, and for non-conserved known non-operon pairs, in *E. coli K12* and *B. subtilis*.




**Supplementary Figure 8: Properties of unsupervised predictions across 124 genomes.** (A) Ability to distinguish same-strand and opposing-strand pairs using comparative features, assessed by Spearman correlation of $p$-values with strandedness, vs. the difference in the proportion of same- and opposing-strand pairs conserved in at least one distant genome. (B) Comparison of our "strand-wise" method for estimating the proportion of same-strand pairs that are in operons, $P(Operon|Same)$, to the *E. coli*-based estimate

$$P(Operon|Same) = .5 \cdot \frac{P_{Genome}(d \in (-20, 30]|Same)}{P_{E.coli}(d \in (-20, 30]|Same)} \qquad \text{(Eq. 1)}$$

proposed by Moreno-Hagelsieb and Collado-Vides (2002). (C) Spearman correlations for our across-cluster gene neighbor score "GNScore" and intergenic distance, versus Spearman correlations for our comparative $p$-values and our distance model (see Sup. Note 1 for a description of GNScore). (D) Contribution of various features to the ability to distinguish same-strand and opposing-strand pairs using comparative features, across 124 genomes. We show the $p$-values from generalized analysis of variance for our logistic regression, using variables in the order shown (GNMinus, GNScore, GNWithin, COG function similarity, and similarity of phylogenetic profiles measured with mutual information; phylogenetic profiles were not actually used for prediction as they do not contain further information). The genomes where the within-cluster score GNWithin was not significantly helpful include 34 genomes (27% of total) which do not have scores because they are in singleton clusters (see Table S3). In B the dashed line indicates $x = y$; in C the dashed line indicates $x = -y$.




**Supplementary Figure 9: Estimates of the proportion of same-strand pairs that are in operons, $P(Operon|Same)$, and of the accuracy of our unsupervised predictions, based on microarrays.** These microarray-based estimates are compared to sequence-based estimates of $P(Operon|Same)$ and to the accuracy implied by the predicted $p$-values. "prior" – $P(Operon|Same)$; "accT" — accuracy of predicted operon pairs; "accF" – accuracy of predicted non-operon pairs.

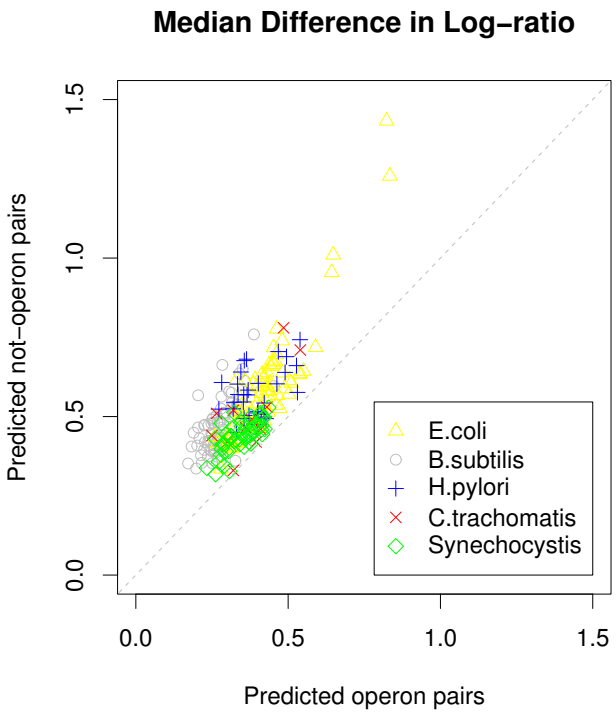
For each genome, we model the distribution of microarray similarity (Pearson $r$) for true operon pairs by using high-confidence predictions ($p > .95$), and model the distribution for non-operon pairs by using opposing-strand pairs. We estimate $P(Operon|Same)$ and both accuracies from a least-squares fit of the density of the relevant set of pairs to a mixture of the two kernel densities for the modeled sets. We use a Gaussian kernel with the default settings in R's *density* function (see http://www.r-project.org/). Because the distribution of $r$ is strongly affected by the extent to which gene levels change, especially for operon pairs, and because genes in operons show greater changes in some species, we split the data into the four quartiles of total change ($\sum \log ratio$ for an arbitrary member of the pair), and compute the density within each quartile. The model density is the average of the within-quartile densities, weighted by the number of pairs in each quartile in the data being fitted. This re-weighting improves the agreement with sequence-based estimates. The accuracies shown for experimentally known pairs are calculated from the observed sensitivity and specificity and the estimated number of operons:

$$P(Operon|p > .5) = \frac{P(Operon|Same) \cdot P(p > .5|Operon)}{P(Operon|Same) \cdot P(p > .5|Operon) + P(\neg Operon|Same) \cdot P(p > .5|\neg Operon)} \quad \text{(Eq. 2)}$$

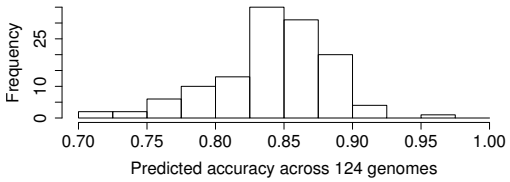
$$P(\neg Operon|p \leq .5) = \frac{P(\neg Operon|Same) \cdot P(p \leq .5|\neg Operon)}{P(Operon|Same) \cdot P(p \leq .5|Operon) + P(\neg Operon|Same) \cdot P(p \leq .5|\neg Operon)} \quad \text{(Eq. 3)}$$

To test the reliability of these estimates, we run them with single microarray conditions removed ("jackknife") or with resampling with replacement (that is, creating a data set of the same size). For the jackknife, we remove entire groups of experiments because microarrays in the same condition are often highly correlated (for conditions see Table S4). 95% confidence intervals from the jackknife are from a $t$ test after multiplying the variance by $\frac{(m-1)\cdot m}{m+1}$, where $m$ is the number of conditions, to reflect the fact that the jackknife estimates are correlated as they mostly use the same data.

The wide variance of the jackknife relative to resampling in all five genomes suggests that operons might be more important to regulation in some conditions than others. The slightly but significantly lower accuracy relative to the sequence-based predictions in *B. subtilis* may indicate a modest violation of either the assumptions used to create our predictions or the assumptions used to estimate accuracy. As discussed in the main text, *Synechocystis* has fewer operons than predicted. The wide confidence interval of the jackknife in *H. pylori* results from having expression data for only two conditions. We do not have the raw data to run the jackknife for *Halobacterium*.




**Supplementary Figure 10: Agreement of unsupervised predictions with each individual microarray experiment.** We show the median absolute difference in normalized log-ratios for predicted operons and non-operon pairs, using only pairs where at least one gene showed some change ($|\log_2(ratio)| \geq 0.3$). The four *E. coli* experiments with high median differences for both sets measured absolute rather than relative levels of mRNA by using genomic DNA as a control.




**Supplementary Figure 11: Estimated accuracy of unsupervised predictions across 124 genomes.** The accuracies are derived from the unsupervised method's $p$-values, using only the genome sequences.

# Supplementary Note 1: Mathematical Description of Methods

| Term | Definition |
|---|---|
| $GNMinus$ | surprisal-subtracted cluster-naive gene neighbor score |
| $GNScore$ | across-cluster gene neighbor score |
| $GNWithin$ | within-cluster gene neighbor score |
| $\vec{X}$ | combined gene neighbor & COG function features |
| $d$ | intergenic distance |
| $CAI$ | codon adaptation index |
| $s_{CAI}$ | similarity of codon adaptation index |
| $Operon$ | same-strand (adjacent) within-operon pairs |
| $\neg Operon$ | same-strand (adjacent) non-operon pairs |
| $Same$ | all same-strand (adjacent) pairs |
| $\neg Same$ | opposing-strand (adjacent) pairs |
| $High$ | same-strand pairs with high values of $P(Same\|\vec{X})$ |
| $Low$ | same-strand pairs with low values of $P(Same\|\vec{X})$ |
| $p_d$ | distance model: $\frac{P(d\|Operon)}{P(d\|Operon)+P(d\|\neg Operon)}$ |
| $p_{CAI}$ | CAI model: $\frac{P(s_{CAI}\|Operon)}{P(s_{CAI}\|Operon)+P(s_{CAI}\|\neg Operon)}$ |
| $Leading$ | a gene is on the leading strand |
| $Lagging$ | a gene is on the lagging strand |
| $Leading_1$ | gene 1 is on the leading strand |
| $Leading_{12}$ | two adjacent genes are on the leading strand |
| $Lagging_{12}$ | two adjacent genes are on the lagging strand |
| $n_{AA}$ | number of times an amino acid occurs in a gene |
| $n_{Codon}$ | number of times a codon occurs in a gene |
| $p_x$ | likelihood ratio for a bin $x$ of a feature |
| $\hat{p_x}$ | maximum likelihood estimate of $p_x$ |

## Features

The gene neighbor scores rely on putative orthologs from bi-directional best BLASTp hits with >75% coverage both ways. We assigned genes to COGs (Tatusov et al. 2001) via reverse position-specific BLAST (Schaffer et al. 2001) against CDD (Marchler-Bauer et al. 2003) as well as using COG membership from NCBI.

To compute our gene neighbor scores, we first cluster the genomes, so that genomes with conserved non-operon pairs will be in the same cluster. We use the fraction of convergently transcribed pairs that are conserved to estimate the distance between two genomes: we use only convergent pairs because a small but significant fraction of divergent pairs are widely conserved, probably because they share a bidirectional promoter. As operons are often shuffled (Itoh et al. 1999), we use a distance cutoff of 5 kilobases between orthologs to define a conserved adjacent pair, rather than requiring that the orthologs are adjacent. We ignore strandedness, as the first step of unsupervised operon prediction requires scores that are not inherently lower for opposing-strand adjacent pairs. We use a somewhat arbitrary threshold of $P(Conserved|Convergent) < 5\%$ to create clusters of similar genomes (any pair of genomes above the threshold in either direction are put into the same cluster).

For each pair of adjacent genes $i$ and $j$ in each genome $G'$ in cluster $C'$, we consider only other genomes $G$ in clusters $C$ which contain predicted orthologs for both $i$ and $j$. Let $f_{G,G'}$ be the fraction of opposing-

strand adjacent pairs in $G$ that are conserved within 5 kb in $G'$. Our scores are:

$$\text{GNScore} = \sum_{C \neq C'} \max_{G \in C} \left( \texttt{if } |\text{position}(i \in G) - \text{position}(j \in G)| < 5\text{kb } \texttt{then } \log \frac{1}{f_{G,G'}} \texttt{ else } 0 \right) \quad \text{(Eq. 4)}$$

$$\text{GNWithin} = \sum_{G \neq G', C = C'} \left( \texttt{if } |\text{position}(i \in G) - \text{position}(j \in G)| < 5\text{kb } \texttt{then } \log \frac{1}{f_{G,G'}} \texttt{ else } -\log \frac{1}{1 - f_{G,G'}} \right) \quad \text{(Eq. 5)}$$

$$\text{GNMinus} = \sum_{G \neq G'} \left( \texttt{if } |\text{position}(i \in G) - \text{position}(j \in G)| < 5\text{kb } \texttt{then } \log \frac{1}{f_{G,G'}} \texttt{ else } -\log \frac{1}{1 - f_{G,G'}} \right) \quad \text{(Eq. 6)}$$

$$\text{GNAll} = \sum_{G \neq G'} \left( \texttt{if } |\text{position}(i \in G) - \text{position}(j \in G)| < 5\text{kb } \texttt{then } \log \frac{1}{f_{G,G'}} \texttt{ else } 0 \right) \quad \text{(Eq. 7)}$$

GNScore is the across-cluster score, GNWithin is the within-cluster score (or 0 if no close relative is in our database), GNMinus is a cluster-naive score which subtracts surprisal values, and GNAll is a naive score. The first three scores were used in the comparative predictions (GNAll is highly dependent on the other three). As shown in Fig. S8D, which assesses the effectiveness of the first step of unsupervised predictions, which is to distinguish same-strand and opposing-strand pairs, the scores contain statistically significant independent information in most genomes. GNMinus is the best single predictor in 77 genomes, GNScore is best in 26 genomes, and GNAll is best in 21 genomes (assessed by single-variable logistic regressions on log likelihood ratios, analogous to the multivariate logistic regression used to make predictions). In *E. coli*, which is within a large gamma-proteobacterial cluster of 26 genomes, $r_{Spearman}(GNScore, Same) = 0.263$ vs. $r_{Spearman}(GNAll, Same) = 0.200$ ($p < 10^{-10}$ from ranked difference correlation).

To analyze pairs of genes by their COG function classes, we assigned pairs to three classes – both genes have the same, meaningful COG function codes (not "R" or "S"); both genes have different yet meaningful COG codes; or one or both genes do not match a characterized COG.

We also experimented with phylogenetic profiles (Pellegrini et al. 1999), otherwise known as phyletic patterns. We first computed clusters of genomes with similar gene content (over 50% of convergent pairs from one genome both having orthologs in the other), analogous to the clustering for the gene neighbor scores. For each pair of adjacent genes, we computed the mutual information between the patterns of co-occurrence of orthologs in the clusters. We converted these scores into log likelihood ratios using smoothed histograms as for other features. Although these log likelihood ratios show a clear correlation with same-strand vs. opposing-strand pairs (median $p$ over 124 genomes of $1.6 \cdot 10^{-17}$), they did not contain further information beyond the gene neighbor scores and COG function class (median $p$ across 124 genomes of 0.48 in generalized analysis of variance, and best $p$ of only 0.004 – see Fig. S8D), so we did not use them to predict operons.

To calculate the codon adaptation index (Sharp and Li, 1987), we need a reference set for each genome. We first compute a codon bias index

$$CBI = \frac{\sum_{Codon} n_{Codon} * log(n_{AA}/n_{Codon})}{\sum_{Codon} n_{Codon}} \quad \text{(Eq. 8)}$$

and then a reference set of 500 COGs which show bias across many genomes. For each genome, we use the most biased 100 genes with at least 300 amino acids among those COGs as our reference set. We then

compute

$$CAI = \exp\left(\frac{\sum_{Codon} n_{Codon} \cdot \log(weight)}{\sum_{Codon} n_{Codon}}\right) \quad \text{(Eq. 9)}$$

where the weight is the reference set's usage of that codon divided by the reference set's usage of the "best" codon for that amino acid. (The "best" codon is the one most frequently used in the reference set.)

To describe the CAI similarity of a pair of genes, we considered

$$s_{CAI} = -(rank(CAI_1) - rank(CAI_2))^2 \quad \text{(Eq. 10)}$$

and

$$s_{CAI} = (rank(CAI_1) - mean(rank)) * (rank(CAI_2) - mean(rank)) \quad \text{(Eq. 11)}$$

Both measures show small, positive, and highly significant ($p < 10^{-8}$) correlations with the ranked scores from our predictions without CAI, in both *E. coli* and *B. subtilis*, so we use the sum of the two scores as our feature. Although in principle, calculating similarity from the full codon usage instead of from codon adaptation (as in Bockhorst et al. 2003a) seems preferable, we did not find highly significant correlations with that method (data not shown).

## Estimating Log Likelihood Ratios for a Feature from a Training Set with Errors

We begin with (i) a training set into two classes, which we will call "1" and "0", (ii) values of the feature for each pair in the training set, and (iii) *a priori* error rates. The first step is to estimate the likelihood ratio for a given bin, or range of values. There is a tradeoff between small bins, which are noisy, and large bins, which assign a single likelihood ratio to values that are substantially different from each other. Our bins for computing likelihood ratios from continuous variables contain 100–200 items (100 items given 2,000 values or less, 200 items per bin given 4,000 values or more, and interpolation in between), and overlapped by 50–100 items. For COG function similarity, we just use the three values (same function class, different function class, or unknown).

To estimate likelihood ratios for a given bin $x$, we define

$$p_x \equiv \frac{p(x|1)}{p(x|1) + p(x|0)} \quad \text{(Eq. 12)}$$

or, equivalently

$$\frac{p_x}{1 - p_x} \equiv \frac{p(x|1)}{p(x|0)} \quad \text{(Eq. 13)}$$

This likelihood ratio is what we wish to estimate. Unlike the more obvious measure $p(1|x)$, $p_x$ does not depend on $p(1)$, the overall probability of being in class "1": we will eventually want to combine several independent likelihood ratios, and furthermore, some of our training sets have a different proportion of "0" and "1" pairs than the proportion of same-strand pairs that are in operons.

Given this bin $x$, we find the maximum likelihood estimate of $p_x$ given a prior

$$\pi(p_x) \propto p_x \cdot (1 - p_x) \quad \text{(Eq. 14)}$$

We need a prior because a uniform prior can lead to estimates below zero or above one, as well as excessive sensitivity to noise. We use this form of prior (a specific Dirichlet) because it is equivalent to adding "pseudocounts" to the observed counts if we have training data without errors (Sjolander et al. 1996), but the choice of parameters (powers for the terms, equivalent to magnitude of the pseudocounts) is *ad hoc* (those parameters are fixed, not estimated from the data). Given a relation between $p(1|x)$ and $p_x$, the

prior $\pi(p_x) \propto p_x \cdot (1 - p_x)$, the observed counts $n_{1x}$ and $n_{0x}$ within the bin, and the binomial likelihood

$$p(n_{1x}, n_{0x}|p_x) \propto p(1|x)^{n_{1x}} \cdot (1 - p(1|x))^{n_{0x}} \quad \text{(Eq. 15)}$$

we can solve for the maximum likelihood value of $p_x$ numerically. In the simplest case of equal numbers of "0" and "1" pairs in training data without errors, $p_x = p(1|x)$, and our prior gives a maximum likelihood estimate for $p_x$ of

$$\hat{p_x} = \frac{n_{1x} + 1}{n_{1x} + n_{0x} + 2} \quad \text{(Eq. 16)}$$

If the training set is not balanced,

$$p(1|x) = \frac{p(1) \cdot p_x}{p(1) \cdot p_x + (1 - p(1)) \cdot (1 - p_x)} \quad \text{(Eq. 17)}$$

where we estimate p(1) from the counts over the entire training set:

$$p(1) = \frac{n_1}{n_1 + n_0} \quad \text{(Eq. 18)}$$

In the most complicated case of estimating distance models from "training" data with both positive and negative errors (based on pairs of genes with high or low ranks from comparative genomics), we have the form

$$p(1|x) = \frac{a + b \cdot p_x}{c + d \cdot p_x} \quad \text{(Eq. 19)}$$

where

$$a = P(High|Same) \cdot P(\neg Operon|High) \quad \text{(Eq. 20)}$$

$$b = P(High|Same) \cdot (P(Operon|High) - P(\neg Operon|High)) \quad \text{(Eq. 21)}$$

$$c = a + P(Low|Same) \cdot P(\neg Operon|Low) \quad \text{(Eq. 22)}$$

$$d = b + P(Low|Same) \cdot (P(Operon|Low) - P(\neg Operon|Low)) \quad \text{(Eq. 23)}$$

Given a likelihood ratio for each bin, we interpolate between the overlapping bins (using ranks of scores, not raw values). We then smooth the results by local regression on the log likelihood ratio $\log(\frac{\hat{p_x}}{1-\hat{p_x}})$ vs. rank($x$) as implemented in *loess* in R (see http://www.r-project.org/). In simulations, smoothing prevents overfitting and increases accuracy to near theoretical limits (data not shown). The span parameter for *loess* is $\alpha = 1$ for the comparative features, as we expect a monotonic distribution and a majority of the gene neighbor scores can be identical (0), and the default setting of $\alpha = .5$ for distance and CAI models.

## Predicting Operons

We first build a model to distinguish same-strand and opposing-strand pairs based on the gene neighbor and COG function features. To combine these highly correlated features, we use logistic regression on the log likelihood ratios (as implemented in R's *glm*, see http://www.r-project.org/) to find the best-fitting linear combination of the log likelihood ratios. This gives us $P(Same|\vec{X})$, the probability that a pair is on the same strand given the comparative features $\vec{X}$. We formalize our key assumption, which is that the distribution of scores for each comparative genomics feature is approximately the same for non-operon pairs (on the same strand) and for opposing-strand pairs, as

$$P(\vec{X}|\neg Operon) \approx P(\vec{X}|\neg Same) \quad \text{(Eq. 24)}$$

and use a prior estimate $P(Operon|Same)$ of the proportion of same-strand pairs that are operon pairs (from our generalization of "directon" counting), giving

$$P(\vec{X}|Same) = P(\vec{X}|Operon) \cdot P(Operon|Same) + P(\vec{X}|\neg Same) \cdot P(\neg Operon|Same) \quad \text{(Eq. 25)}$$

$$\frac{P(\vec{X}|Operon)}{P(\vec{X}|\neg Operon)} = \frac{\frac{P(\neg Same)}{P(Same)} \cdot \frac{P(Same|\vec{X})}{P(\neg Same|\vec{X})} - P(\neg Operon|Same)}{P(Operon|Same)} \quad \text{(Eq. 26)}$$

We use the output of the logistic regression to split the same-strand pairs into likely operon (high-probability) and non-operon (low-probability) pairs. We use the ranked scores and a cutoff chosen to give approximately the correct number of predicted operons:

$$percentile\left(P(Same|\vec{X})\right) > P(\neg Operon|Same) \quad \text{(Eq. 27)}$$

We estimate the accuracy on both sides of this split from the number of same-strand pairs with low $P(\vec{X}|Operon)$, the number of opposite-strand pairs with high $P(\vec{X}|Operon)$, and the prior:

$$P(\neg Operon|High) = \frac{P(High|\neg Same)}{P(High|Same)} \cdot P(\neg Operon|Same) \quad \text{(Eq. 28)}$$

$$P(Operon|Low) = \frac{P(Operon|Same) - P(Operon|High) \cdot P(High|Same)}{P(Low|Same)} \quad \text{(Eq. 29)}$$

In practice, the accuracy of the split is modest but sufficient: in *E. coli*, the split has 78.9% accuracy on known operons and 74.2% accuracy on known non-operons, vs. predicted 85.0% and 64.0%. In *B. subtilis*, the accuracy is 88.0% and 50.0% vs. predicted 79.1% and 73.9%.

We produce a distance model

$$\frac{p_d}{1-p_d} \equiv \frac{P(d|Operon)}{P(d|\neg Operon)} \quad \text{(Eq. 30)}$$

from this "training" set. Intuitively, we wish to subtract some fraction of $P(d|Low)$ from $P(d|High)$, and divide by the proportion subtracted out, because errors in our preliminary predictions will reduce the difference between the two distributions. Actually doing a subtraction is undesirable because it can lead to illegal estimates such as $p_d < 0$ or $p_d > 1$, even after smoothing. Instead, we produce a maximum likelihood estimate of $p_d$ for each bin from our accuracy estimates – which imply a relation between $p_d$ and $\frac{P(High|d)}{P(Low|d)}$ – and a prior, and smooth the results. Similarly, we use the combination of the distance model and the logistic regression to estimate a $p$-value for our CAI-derived score, but without the "subtraction" step, as we no longer have an estimate of accuracy. (We introduce CAI here, instead of in the logistic regression, because it shows strong strand bias in some genomes, and thus would violate our assumption.) Our final predicted within-operon pairs are those where

$$\frac{P(Operon|\vec{X}, d, s_{CAI})}{P(\neg Operon|\vec{X}, d, s_{CAI})} = \frac{P(Operon|Same)}{P(\neg Operon|Same)} \frac{P(\vec{X}|Operon)}{P(\vec{X}|\neg Operon)} \frac{p_d}{1-p_d} \frac{p_{CAI}}{1-p_{CAI}} > \frac{1}{2} \quad \text{(Eq. 31)}$$

For our supervised predictions, we use smoothed histograms on the known operon and non-operon pairs to generate likelihood ratios for each feature, logistic regression to combine the gene neighbor and COG scores, and the standard formula

$$\frac{P(\vec{x}|Operon)}{P(\vec{x}|\neg Operon)} = \prod_i \frac{P(x_i|Operon)}{P(x_i|\neg Operon)} \quad \text{(Eq. 32)}$$

to combine the results of the logistic regression with the distance and CAI similarity scores. We make predictions for pairs from the training set with 100-fold cross-validation, and make predictions for other pairs with the mean of the log likelihood ratios from these 100 models.

*Assumption of conditional independence.* Both the supervised and unsupervised methods use Bayes' rule to combine likelihood ratios from the gene neighbor method and from the distance model. This implicitly assumes that the two features are conditionally independent (that is, that they are independent once we know whether pair is within an operon or not). This is a standard machine learning assumption, known as "naive Bayes," and it is biologically plausible. We have tested the conditional independence assumption on known operon and not-operon pairs from E. coli. There is no correlation between GNMinus and distance within known pairs (Spearman $r = -0.06$, $p = 0.13$). There is a significant negative correlation within the putative not-operon pairs (Spearman $r$ = -0.26, $p = 2.5 \cdot 10^{-9}$), which we suspect is partly due to the presence of operon pairs within that set. The correlation between GNMinus and distance within the predicted non-operon set from our unsupervised method is weak (Spearman $r = -0.07$, $p = 0.005$).

The fact that the distance model is built from the gene neighbor method-based predictions does not affect whether the likelihood ratios are conditionally independent. In fact, the mathematical derivation given above of the process for building a genome-specific distance model assumes conditional independence. As discussed in a previous section of this note, smoothing while estimating log-likelihood ratios prevents overfitting, which might otherwise lead to a violation of conditional independence.

## Estimating the Number of Operons

For genomes without strand bias, the number of transcription unit changes is twice the number of direction changes (Ermolaeva et al. 2001; Cherry 2003), or, in our notation,

$$1 - P(Operon|Same) \cdot P(Same) = 2 \cdot P(\neg Same) \quad \text{(Eq. 33)}$$

which gives

$$P(Operon|Same) = 2 - \frac{1}{P(Same)} \quad \text{(Eq. 34)}$$

which can equivalently be derived from the neutral assumption

$$P(Same|\neg Operon) = \frac{1}{2} \quad \text{(Eq. 35)}$$

Our strand-naive model assumes that transcripts are randomly assorted between strands in a biased manner:

$$P(Operon|Leading_1) = P(Operon|Lagging_1) \quad \text{(Eq. 36)}$$

Given a first gene on (say) the leading strand, the probability that the next gene is on the same strand is

$$P(Leading_2|Leading_1) = P(Operon|Leading_1) + P(\neg Operon|Leading_1) \cdot P(Leading) \quad \text{(Eq. 37)}$$

where $P(Leading)$ is the extent of strand bias. This gives

$$P(Operon|Same) = \frac{1 - \frac{P(Leading)^2 + P(Lagging)^2)}{P(Same)}}{1 - (P(Leading)^2 + P(Lagging)^2)} \quad \text{(Eq. 38)}$$

This "strand-neutral" model implies that pairs of genes on the lagging strand are more likely to be in operons, as adjacent transcripts on the lagging strand are less likely to arise by chance:

$$P(Operon|Leading_2, Leading_1) = \frac{P(Operon|Leading_1)}{P(Leading_2|Leading_1)}$$
$$< \frac{P(Operon|Lagging_1)}{P(Lagging_2|Lagging_1)} = P(Operon|Lagging_2, Lagging_1) \quad \text{(Eq. 39)}$$

because this model assumes that the numerators are equal.
Our "strand-wise model" assumes that

$$P(Operon|Same, Leading_{12}) = P(Operon|Same, Lagging_{12}) = P(Operon|Same) \quad \text{(Eq. 40)}$$

By the same reasoning as above this gives

$$P(Operon|Leading_1) > P(Operon|Lagging_1) \quad \text{(Eq. 41)}$$

and we derive

$$P(Operon|Same) = \frac{P(Operon|Lagging_1)}{P(Lagging_2|Lagging_1)} \quad \text{(Eq. 42)}$$

$$a \cdot P(Operon|Lagging_1)^2 + b \cdot P(Operon|Lagging_1) + c = 0 \quad \text{(Eq. 43)}$$

$$a = \frac{P(Leading_2|Leading_1)}{P(Lagging_2|Lagging_1)} \quad \text{(Eq. 44)}$$

$$b = -2 \cdot P(Leading_2|Leading_1) \quad \text{(Eq. 45)}$$

$$c = P(Leading_2|Leading_1) + P(Lagging_2|Lagging_1) - 1 \quad \text{(Eq. 46)}$$

**Additional References:**

Marchler-Bauer, A., Anderson, J.B., DeWeese-Scott, C., Fedorova, N.D., Geer, L.Y., He, S., Hurwitz, D.I., Jackson, J.D., Jacobs, A.R., Lanczycki, C.J., Liebert, C.A., Liu, C., Madej, T., Marchler, G.H., Mazumder, R., Nikolskaya, A.N., Panchenko, A.R., Rao, B.S., Shoemaker, B.A., Simonyan, V., Song, J.S., Thiessen, P.A., Vasudevan, S., Wang, Y., Yamashita, R.A., Yin, J.J., and Bryant, S.H. CDD: a curated Entrez database of conserved domain alignments. *Nucleic Acids Res.* **31**:383–7.

Schaffer, A.A., Aravind, L., Madden, T.L., Shavirin, S., Spouge, J.L., Wolf, Y.I., Koonin, E.V., and Altschul, S.F. Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. *Nucleic Acids Res.* **29**:2994–3005.

Sjolander, K., Karplus, K., Brown, M., Hughey, R., Krogh, A., Mian, I.S., and Haussler,D. (1996) Dirichlet Mixtures: A Method for Improving Detection of Weak but Significant Protein Sequence Homology. *Comput. Appl. Biosci.* 12(4): 327–45.

## Supplementary Note 2: Gene Starts and Potential Pseudogenes in *B. anthracis str. Ames*

To investigate the underlying causes of the unusual distance model for *B. anthracis str. Ames*, we examined both potential pseudogenes – intergenic regions with long stretches homologous to annotated ORFs – and

potential gene truncations – intergenic regions homologous to the *B. cereus* ortholog of one of the adjacent genes. These "pseudogenes" and "truncations" could reflect sequencing errors, errors in gene identification, errors in gene start prediction, ORFs that remain functional despite frameshifts or premature stop codons, as well as genuine disruption or truncation of a functioning ORF. The effect on our distance model, however, will be the same in all cases.

We estimate that 7% of same-strand pairs in *B. anthracis* have intergenic distances that are affected by one or both phenomena (259/3749), leading to significant numbers of conserved pairs with large gaps. In theory, this should increase our distance model $p_d$ by 0.07 at large distances. The minimum value of our distance model $p_d$ (which is attained at the highest distances) is .114 in *B. anthracis*, compared to .042 in *B. subtilis* and .025 in *E. coli K12*. Thus, these truncations, pseudogenes, and erroneous start predictions are a major factor in the differences in distance models.

We identified the best hit of each intergenic region to coding regions in our database of 124 genomes, using BLASTn and an E-value cutoff of $10^{-5}$. We examined putative pseudogenes by considering only matches over 200 bases long, giving 175 pseudogenes in *B. anthracis*, 9 in *B. subtilis*, 89 in *E. coli K12*, and 12 in *Synechocystis*. (Note that this analysis does not consider the possibility that some annotated ORFs are pseudogenes, as that would not affect our distance models.) All of these potential pseudogenes were highly similar to annotated ORFs (at least 78% identity), and many matches were very long (median length in *B. anthracis* of 733 bases, vs. 279 for the same set in *E. coli*). Most of the *B. anthracis* pseudogenes had their best hits to *B. cereus* ORFs (159/175).

We identified potential pseudogenes within ancestral operons by selecting putative pseudogenes between same-strand pairs that were conserved near each other in a distant genome ($GNScore > 0$). This gave 12/108 same-strand pairs in *B. anthracis*, versus 0/4 in *B. subtilis*, 2/51 in *E. coli*, and 0/8 in *Synechocystis*. For comparison, 24% of all same-strand pairs in *B. anthracis* are conserved in a distant genome (versus 30% in *B. subtilis*, 36% in *E. coli*, and 14% in *Synechocystis*).

The two most conserved *B. anthracis* operons containing putative pseudogenes both include a frame-shifted, highly similar, positional homolog from *B. cereus* (>85% nucleotide identity). The pseudogenes are not due to sequencing errors, as shown by comparing to the sequences for other strains at GenBank (e.g. 100% identity in strain A2012), and not annotated in the GenBank entry for the complete *B. anthracis str. Ames* genome (NC_003997.3). The first pair is annotated in *B. anthracis* as an ABC transporter permease and iron compound-binding protein, with the pseudogene's homolog in *B. cereus* annotated as 2-aminoethylphosphonate transport ATP-binding protein, and the second pair is annotated in *B. anthracis* as flagellar hook-associated protein and flagellar hook-associated FliD, with the pseudogene's homolog in *B. cereus* annotated as Flagellar hook-associated protein 3.

We also identified potential gene truncations or errors in predicted gene starts in *B. anthracis*, by analyzing the best hits (of any length) between same-strand pairs in *B. anthracis* that both had close (>70% identity) homologs in *B. cereus* that were adjacent or had only one intervening gene. (We did not use best hits because if a gene is severely truncated, the true ortholog might not have the highest BLAST score). This left 192 same-strand pairs containing intergenic homology to an ORF (in any genome), of which 166 were syntenic (the best hit of the intergenic region was to the adjacent or intervening ORF in *B. cereus*). Of these, 87 were apparent truncations (only three of these were over 200 bases long), and 79 were apparent pseudogenes (58 of these were over 200 bases long). For the apparent gene truncations, the homology was most often to the downstream gene. This suggests that problems with the gene start predictions, rather than truncation of an ORF by the introduction of a stop codon, might be the underlying cause. Almost all of the "truncations" and most of the "pseudogenes" were gap-free alignments to ORFs.

Our estimate of 259 same-strand pairs with intergenic regions disrupted in *B. anthracis* is based on the "success rate" of 166/192 from the *B. cereus* synteny analysis multiplied by the total number of intergenic regions with homology to ORFs that are between same-strand pairs (300).

# Supplementary Tables

**Supplementary Table 1: Agreement of our unsupervised and supervised predictions with experimentally identified operon and non-operon pairs in *E. coli* and *B. subtilis*.** AOC is the area under the operating curve shown in Figure 3, or the probability that an operon pair will have a better score than a non-operon pair if both pairs are chosen at random. Default sensitivity (fraction of known operon pairs which are correctly predicted) and specificity (fraction of known non-operon pairs which are correctly predicted) are computed with a threshold of predicted $p > 0.5$, and maximum accuracy is the maximum over all possible thresholds of the average of sensitivity and specificity. The unsupervised microarray-based predictions, which are shown only in this table, use a logistic regression of the microarray data (rank of Pearson $r$, total intensity, and total absolute change of the pair, with pairwise interactions) versus the usual unsupervised predictions (thresholded at 0.5).

For comparison, we show results from our supervised predictions, from Salgado et al. 2000 for *E. coli* (using distance and Monica Riley's functional classification, or just distance), from Sabatti et al. 2002 for *E. coli* (using correlation in microarray data and/or distance as features, on a somewhat different training set), from Bockhorst et al. 2003b for *E. coli* (distance-only or distance plus microarrays and further sequence-based features), from Moreno-Hagelsieb and Collado-Vides 2002 for *B. subtilis* (using a distance model trained in *E. coli*), and from De Hoon et al. 2004 for *B. subtilis* (using distance and/or microarray correlation, and a much larger unpublished training set). We do not show the results of Bockhorst et al. 2003a because they report accuracy for predicting transcripts, not individual pairs of genes.

| Measure | AOC | Max. Acc. | Def. Sens. | Def. Spec. |
|---|---|---|---|---|
| **E. coli** | | | | |
| Unsupervised (Sequence-only) | 0.920 | 0.852 | 0.883 | 0.799 |
| Distance-only | 0.886 | 0.829 | 0.794 | 0.857 |
| Unsupervised with microarrays | 0.925 | 0.863 | 0.890 | 0.817 |
| Microarray-only | 0.820 | 0.750 | 0.834 | 0.660 |
| Supervised (Sequence-only) | 0.919 | 0.859 | 0.865 | 0.850 |
| Salgado et al. 2000 | – | 0.87 | – | – |
| Distance-only | – | 0.82 | – | – |
| Sabatti et al. 2002 | – | 0.88 | 0.88 | 0.88 |
| Distance-only | – | 0.83 | 0.84 | 0.82 |
| Microarray-only | – | 0.76 | 0.82 | 0.70 |
| Bockhorst et al. 2003b | 0.929 | – | 0.78 | 0.90 |
| Distance-only | 0.915 | – | – | – |
| **B. subtilis** | | | | |
| Unsupervised (Sequence-only) | 0.888 | 0.815 | 0.909 | 0.710 |
| Distance-only | 0.882 | 0.863 | 0.825 | 0.863 |
| Unsupervised with microarrays | 0.885 | 0.844 | 0.922 | 0.727 |
| Microarray-only | 0.748 | 0.692 | 0.804 | 0.545 |
| Supervised (Sequence-only) | 0.907 | 0.868 | 0.877 | 0.847 |
| Moreno-Hagelsieb & Collado-Vides 2002 | – | 0.82 | – | – |
| De Hoon et al. 2004 | – | 0.884 | 0.888 | 0.879 |
| Distance-only | – | 0.856 | 0.821 | 0.890 |
| Microarray-only | – | 0.796 | 0.801 | 0.791 |

**Supplementary Table 2: Agreement of predictions, based on various combinations of features or distance models, with microarray data.** "*E. coli* $p_d$" means using the supervised *E. coli* distance model (computed using all of the training pairs and shown in Fig. 3), either by itself or combined with comparative predictions. "GNMinus" is a gene neighbor score with surprisal subtraction (see Sup. Note 1). "Relaxed assumption" predictions are computed by assuming that 10% of same-strand non-operon pairs look like operon pairs, which in practice tends to downplay the comparative features and put a higher weight on the distance model. A potential biological basis for

this assumption is that ancestral operons might split into separate transcripts without being shuffled, but no overall benefit is seen.

$r_S(m, r_\mu)$ is the Spearman correlation of each measure $m$ with the similarity of expression (Pearson $r$ in microarray data). This indicates whether the measure ranks pairs of genes appropriately. $r_S(m > .5, r_\mu)$ is the Spearman correlation coefficient on a thresholded measure (only considering whether $m$ is above 0.5 or not). This indicates whether the default threshold is effective. $r_S(m, p_{Operon})$ is the Spearman correlation with unsupervised predictions using the full set of the features. $P(Operon|Same)$ is the proportion of same-strand pairs that are in operons. $P(Same)$ is the proportion of adjacent pairs that are on the same strand. $P(Leading)$ is the coding strand bias, or the proportion of genes estimated to be on the leading strand. $n_{pairs}$ is the number of same-strand pairs for which a value of expression similarity (Pearson $r$) is available.

*E. coli K12*: $P(Operon|Same) = 0.57$, $P(Same) = 0.70$, $P(Leading) = 0.56$, $n_{pairs} = 2838/3005$

| **Measure ($m$)** | $r_S(m, r_\mu)$ | $r_S(m > .5, r_\mu)$ | $r_S(m, p_{Operon})$ |
|---|---|---|---|
| Untrained | 0.494 | 0.421 | 1.000 |
| Logistic regression | 0.406 | 0.345 | 0.768 |
| GNMinus | 0.375 | – | 0.702 |
| Distance-only | 0.401 | 0.394 | 0.873 |
| E.coli $p_d$ | 0.387 | 0.396 | 0.837 |
| Raw Sep. | 0.384 | – | 0.808 |
| without CAI | 0.489 | 0.422 | 0.993 |
| E.coli $p_d$ | 0.497 | 0.429 | 0.977 |
| relaxed assumption | 0.483 | 0.418 | 0.995 |

*B. subtilis*: $P(Operon|Same) = 0.52$, $P(Same) = 0.73$, $P(Leading) = 0.74$, $n_{pairs} = 2278/3020$

| **Measure ($m$)** | $r_S(m, r_\mu)$ | $r_S(m > .5, r_\mu)$ | $r_S(m, p_{Operon})$ |
|---|---|---|---|
| Untrained | 0.461 | 0.392 | 1.000 |
| Logistic regression | 0.335 | 0.284 | 0.840 |
| GNMinus | 0.315 | – | 0.775 |
| Distance-only | 0.420 | 0.429 | 0.733 |
| E.coli $p_d$ | 0.423 | 0.427 | 0.729 |
| Raw Sep. | 0.423 | – | 0.727 |
| without CAI | 0.456 | 0.386 | 0.993 |
| E.coli $p_d$ | 0.471 | 0.434 | 0.981 |
| relaxed assumption | 0.470 | 0.397 | 0.994 |

*H. pylori*: $P(Operon|Same) = 0.70$, $P(Same) = 0.78$, $P(Leading) = 0.60$, $n_{pairs} = 1194/1226$

| **Measure ($m$)** | $r_S(m, r_\mu)$ | $r_S(m > .5, r_\mu)$ | $r_S(m, p_{Operon})$ |
|---|---|---|---|
| Untrained | 0.343 | 0.308 | 1.000 |
| Logistic regression | 0.231 | – | 0.669 |
| GNMinus | 0.177 | – | 0.426 |
| Distance-only | 0.275 | 0.328 | 0.882 |
| E.coli $p_d$ | 0.271 | 0.307 | 0.746 |
| Raw Sep. | 0.270 | – | 0.704 |
| without CAI | 0.345 | 0.312 | 0.993 |
| E.coli $p_d$ | 0.349 | 0.262 | 0.938 |
| relaxed assumption | 0.338 | 0.267 | 0.998 |

*C. trachomatis*: $P(Operon|Same) = 0.60$, $P(Same) = 0.72$, $P(Leading) = 0.57$, $n_{pairs} = 615/641$

| **Measure** ($m$) | $r_S(m, r_\mu)$ | $r_S(m > .5, r_\mu)$ | $r_S(m, p_{Operon})$ |
|---|---|---|---|
| Untrained | 0.303 | 0.289 | 1.000 |
| Logistic regression | 0.167 | 0.0965 | 0.804 |
| GNMinus | 0.192 | – | 0.751 |
| Distance-only | 0.260 | 0.309 | 0.750 |
| E.coli $p_d$ | 0.278 | 0.310 | 0.740 |
| Raw Sep. | 0.267 | – | 0.740 |
| without CAI | 0.305 | 0.292 | 0.998 |
| E.coli $p_d$ | 0.309 | 0.283 | 0.981 |
| relaxed assumption | 0.305 | 0.315 | 0.995 |

*Synechocystis PCC 6803*: $P(Operon|Same) = 0.48$, $P(Same) = 0.66$, $P(Leading) = 0.53$, $n_{pairs} = 1947/2093$

| **Measure** ($m$) | $r_S(m, r_\mu)$ | $r_S(m > .5, r_\mu)$ | $r_S(m, p_{Operon})$ |
|---|---|---|---|
| Untrained | 0.268 | 0.232 | 1.000 |
| Logistic regression | 0.222 | 0.233 | 0.697 |
| GNMinus | 0.218 | – | 0.626 |
| Distance-only | 0.159 | 0.116 | 0.716 |
| E.coli $p_d$ | 0.170 | 0.118 | 0.629 |
| Raw Sep. | 0.171 | – | 0.631 |
| without CAI | 0.267 | 0.232 | 0.985 |
| E.coli $p_d$ | 0.259 | 0.225 | 0.849 |
| relaxed assumption | 0.265 | 0.240 | 0.997 |

*Halobacterium NRC-1*: $P(Operon|Same) = 0.28$, $P(Same) = 0.58$, $P(Leading) = 0.53$, $n_{pairs} = 1114/1211$

| **Measure** ($m$) | $r_S(m, r_\mu)$ | $r_S(m > .5, r_\mu)$ | $r_S(m, p_{Operon})$ |
|---|---|---|---|
| Untrained | 0.215 | 0.190 | 1.000 |
| Logistic regression | 0.155 | 0.208 | 0.672 |
| GNMinus | 0.150 | – | 0.332 |
| Distance-only | 0.198 | 0.210 | 0.817 |
| E.coli $p_d$ | 0.189 | 0.127 | 0.729 |
| Raw Sep. | 0.191 | – | 0.720 |
| without CAI | 0.221 | 0.181 | 0.977 |
| E.coli $p_d$ | 0.219 | 0.226 | 0.913 |
| relaxed assumption | 0.215 | 0.179 | 0.996 |

**Supplementary Table 3: Genomes included in the analyses.** We show the name of the genome; its NCBI taxonomy identifier; the source from which we obtained the genome sequence; the clustering of genomes into related groups with conserved gene order, showing the the genus of a member of the cluster and the cluster size, if the genome was placed in a cluster, or a dash if the genome was in a cluster by itself; the proportion $P(Same)$ of adjacent pairs that are on the same strand; the strand-wise estimate $P(Operon|Same)$ of the proportion of same-strand pairs that are in operons; and the number of same-strand adjacent pairs. The source includes a "*" if we generated gene models ourselves with CRITICA (Badger and Olsen 1999) rather than obtaining them with the sequence. NCBI – http://www.ncbi.nlm.nih.gov/genomes/MICROBES/Complete.html; TIGR – http://www.tigr.org; JGI – http://www.jgi.doe.gov/.

| Name | TaxId | Source | Cluster | $P(Same)$ | $P(Operon\|Same)$ | $n_{Same}$ |
|---|---|---|---|---|---|---|
| Aeropyrum pernix | 56636 | NCBI | – | 0.568 | 0.233 | 1046 |
| Agrobacterium tumefaciens C58 | 181661 | NCBI | Bruce7 | 0.694 | 0.548 | 3159 |
| Agrobacterium tumefaciens C58 (UW) | 180835 | NCBI | Bruce7 | 0.678 | 0.514 | 3158 |
| Archaeoglobus fulgidus | 2234 | NCBI | – | 0.705 | 0.577 | 1706 |
| Bacillus anthracis str. Ames | 198094 | NCBI | Bacil9 | 0.706 | 0.417 | 3749 |
| Bacillus cereus ATCC 14579 | 226900 | NCBI | Bacil9 | 0.713 | 0.431 | 3734 |
| Bacillus halodurans | 86665 | NCBI | Bacil9 | 0.757 | 0.533 | 3076 |
| Bacillus subtilis | 1423 | NCBI | Bacil9 | 0.734 | 0.517 | 3020 |
| Bacteroides thetaiotaomicron VPI-5482 | 226186 | NCBI | – | 0.758 | 0.672 | 3622 |
| Bifidobacterium longum NCC2705 | 206672 | NCBI | – | 0.691 | 0.506 | 1195 |
| Borrelia burgdorferi B31 | 224326 | NCBI | – | 0.729 | 0.581 | 620 |
| Bradyrhizobium japonicum | 375 | NCBI | Bruce7 | 0.672 | 0.502 | 5593 |
| Brucella melitensis | 29459 | NCBI | Bruce7 | 0.696 | 0.542 | 2226 |
| Brucella suis 1330 | 204722 | NCBI | Bruce7 | 0.670 | 0.489 | 2188 |
| Buchnera aphidicola str. APS | 107806 | NCBI | Wiggl19 | 0.741 | 0.626 | 418 |
| Buchnera aphidicola str. Bp | 224915 | NCBI | Wiggl19 | 0.722 | 0.584 | 364 |
| Buchnera aphidicola str. Sg | 198804 | NCBI | Wiggl19 | 0.721 | 0.581 | 393 |
| Campylobacter jejuni | 197 | NCBI | – | 0.789 | 0.717 | 1290 |
| Caulobacter crescentus CB15 | 190650 | NCBI | – | 0.654 | 0.461 | 2445 |
| Chlamydia muridarum | 83560 | NCBI | Chlam6 | 0.706 | 0.575 | 638 |
| Chlamydia trachomatis | 813 | NCBI | Chlam6 | 0.716 | 0.597 | 641 |
| Chlamydophila caviae GPIC | 227941 | NCBI | Chlam6 | 0.703 | 0.566 | 702 |
| Chlamydophila pneumoniae AR39 | 115711 | NCBI | Chlam6 | 0.725 | 0.610 | 806 |
| Chlamydophila pneumoniae CWL029 | 115713 | NCBI | Chlam6 | 0.717 | 0.597 | 756 |
| Chlamydophila pneumoniae J138 | 138677 | NCBI | Chlam6 | 0.721 | 0.605 | 771 |
| Chlorobium tepidum TLS | 194439 | NCBI | – | 0.676 | 0.507 | 1522 |
| Clostridium acetobutylicum | 1488 | NCBI | Clost3 | 0.774 | 0.559 | 2842 |
| Clostridium perfringens | 1502 | NCBI | Clost3 | 0.770 | 0.421 | 2048 |
| Clostridium tetani E88 | 212717 | NCBI | Clost3 | 0.776 | 0.517 | 1841 |
| Corynebacterium efficiens YS-314 | 196164 | NCBI | Mycob5 | 0.645 | 0.421 | 1902 |
| Corynebacterium glutamicum | 196627 | NCBI | Mycob5 | 0.674 | 0.497 | 2017 |
| Coxiella burnetii RSA 493 | 227377 | NCBI | – | 0.678 | 0.500 | 1363 |
| Deinococcus radiodurans | 1299 | NCBI | – | 0.665 | 0.489 | 1993 |
| Desulfovibrio vulgaris | 881 | TIGR | – | 0.689 | 0.546 | 2431 |
| Enterococcus faecalis V583 | 226185 | NCBI | – | 0.790 | 0.580 | 2459 |
| Escherichia coli CFT073 | 199310 | NCBI | Wiggl19 | 0.646 | 0.445 | 3475 |
| Escherichia coli K12 | 83333 | NCBI | Wiggl19 | 0.702 | 0.570 | 3005 |
| Escherichia coli O157:H7 | 83334 | NCBI | Wiggl19 | 0.735 | 0.628 | 3941 |
| Escherichia coli O157:H7 EDL933 | 155864 | NCBI | Wiggl19 | 0.727 | 0.613 | 3868 |
| Fusobacterium nucleatum, ATCC25586 | 190304 | NCBI | – | 0.806 | 0.754 | 1667 |
| Haemophilus influenzae Rd KW20 | 71421 | NCBI | – | 0.733 | 0.628 | 1256 |
| Halobacterium sp. NRC-1 | 64091 | NCBI | – | 0.584 | 0.284 | 1211 |
| Helicobacter pylori 26695 | 85962 | NCBI | Helic2 | 0.778 | 0.705 | 1226 |
| Helicobacter pylori J99 | 85963 | NCBI | Helic2 | 0.780 | 0.712 | 1163 |
| Lactobacillus plantarum WCFS1 | 220668 | NCBI | – | 0.740 | 0.541 | 2227 |
| Lactococcus lactis subsp. lactis | 1360 | NCBI | – | 0.808 | 0.632 | 1831 |
| Leptospira interrogans, 56601 | 189518 | NCBI | – | 0.656 | 0.459 | 3099 |
| Listeria innocua | 1642 | NCBI | Liste2 | 0.805 | 0.632 | 2388 |
| Listeria monocytogenes EGD-e | 169963 | NCBI | Liste2 | 0.788 | 0.606 | 2244 |
| Mesorhizobium loti | 381 | NCBI | Bruce7 | 0.662 | 0.483 | 4464 |
| Methanocaldococcus jannaschii | 2190 | NCBI | – | 0.707 | 0.580 | 1223 |
| Methanopyrus kandleri AV19 | 190192 | NCBI | – | 0.663 | 0.482 | 1119 |
| Methanosarcina acetivorans C2A | 188937 | NCBI | Metha2 | 0.667 | 0.500 | 3030 |
| Methanosarcina mazei Goe1 | 192952 | NCBI | Metha2 | 0.683 | 0.532 | 2303 |
| Methanothermobacter thermautotroph. | 187420 | NCBI | – | 0.748 | 0.655 | 1401 |
| Mycobacterium leprae | 1769 | NCBI | Mycob5 | 0.688 | 0.484 | 1105 |
| Mycobacterium tuberculosis CDC1551 | 83331 | NCBI | Mycob5 | 0.648 | 0.435 | 2715 |
| Mycobacterium tuberculosis H37Rv | 83332 | NCBI | Mycob5 | 0.668 | 0.484 | 2625 |
| Mycoplasma gallisepticum R | 233150 | NCBI | Mycop3 | 0.835 | 0.701 | 606 |
| Mycoplasma genitalium | 2097 | NCBI | Mycop3 | 0.839 | 0.705 | 406 |
| Mycoplasma penetrans | 28227 | NCBI | – | 0.801 | 0.591 | 831 |
| Mycoplasma pneumoniae | 2104 | NCBI | Mycop3 | 0.820 | 0.678 | 565 |
| Mycoplasma pulmonis | 2107 | NCBI | – | 0.757 | 0.650 | 592 |
| Neisseria meningitidis MC58 | 122586 | NCBI | Neiss2 | 0.736 | 0.636 | 1531 |
| Neisseria meningitidis Z2491 | 122587 | NCBI | Neiss2 | 0.737 | 0.638 | 1521 |
| Nitrosomonas europaea ATCC 19718 | 228410 | NCBI | – | 0.746 | 0.652 | 1837 |
| Nostoc sp. PCC 7120 | 103690 | NCBI | – | 0.634 | 0.420 | 3402 |

| Name | TaxId | Source | Cluster | $P(Same)$ | $P(Operon\|Same)$ | $n_{Same}$ |
|---|---|---|---|---|---|---|
| Oceanobacillus iheyensis | 182710 | NCBI | Bacil9 | 0.752 | 0.551 | 2632 |
| Pasteurella multocida | 747 | NCBI | – | 0.727 | 0.613 | 1465 |
| Pseudomonas aeruginosa PAO1 | 208964 | NCBI | Pseud3 | 0.706 | 0.575 | 3929 |
| Pseudomonas putida KT2440 | 160488 | NCBI | Pseud3 | 0.684 | 0.533 | 3662 |
| Pseudomonas syringae | 223283 | NCBI | Pseud3 | 0.691 | 0.550 | 3783 |
| Pyrobaculum aerophilum | 13773 | NCBI | – | 0.634 | 0.417 | 1651 |
| Pyrococcus abyssi | 29292 | NCBI | Pyroc3 | 0.689 | 0.540 | 1219 |
| Pyrococcus furiosus DSM 3638 | 186497 | NCBI | Pyroc3 | 0.690 | 0.545 | 1425 |
| Pyrococcus horikoshii | 53953 | NCBI | Pyroc3 | 0.635 | 0.398 | 1143 |
| Ralstonia solanacearum | 305 | NCBI | – | 0.728 | 0.609 | 2506 |
| Rickettsia conorii | 781 | NCBI | Ricke2 | 0.769 | 0.679 | 1056 |
| Rickettsia prowazekii | 782 | NCBI | Ricke2 | 0.751 | 0.648 | 627 |
| Salmonella enterica, Typhi | 90370 | NCBI | Wiggl19 | 0.721 | 0.597 | 3169 |
| Salmonella enterica, Typhi Ty2 | 209261 | NCBI | Wiggl19 | 0.720 | 0.596 | 3111 |
| Salmonella typhimurium LT2 | 99287 | NCBI | Wiggl19 | 0.708 | 0.573 | 3153 |
| Shewanella oneidensis MR-1 | 211586 | NCBI | Wiggl19 | 0.685 | 0.530 | 2962 |
| Shigella flexneri 2a str. 2457T | 198215 | NCBI | Wiggl19 | 0.692 | 0.548 | 2816 |
| Shigella flexneri 2a str. 301 | 198214 | NCBI | Wiggl19 | 0.691 | 0.548 | 2888 |
| Sinorhizobium meliloti | 382 | NCBI | Bruce7 | 0.687 | 0.537 | 2295 |
| Staphylococcus aureus, Mu50 | 158878 | NCBI | Bacil9 | 0.755 | 0.573 | 2048 |
| Staphylococcus aureus, MW2 | 196620 | NCBI | Bacil9 | 0.752 | 0.543 | 1980 |
| Staphylococcus aureus, N315 | 158879 | NCBI | Bacil9 | 0.747 | 0.548 | 1938 |
| Staphylococcus epidermidis | 176280 | NCBI | Bacil9 | 0.721 | 0.481 | 1743 |
| Streptococcus agalactiae 2603V/R | 208435 | NCBI | Strep7 | 0.812 | 0.647 | 1724 |
| Streptococcus agalactiae NEM316 | 211110 | NCBI | Strep7 | 0.827 | 0.669 | 1732 |
| Streptococcus mutans UA159 | 210007 | NCBI | Strep7 | 0.806 | 0.632 | 1580 |
| Streptococcus pneumoniae R6 | 171101 | NCBI | Strep2 | 0.808 | 0.657 | 1651 |
| Streptococcus pneumoniae TIGR4 | 170187 | NCBI | Strep2 | 0.799 | 0.614 | 1674 |
| Streptococcus pyogenes M1 GAS | 160490 | NCBI | Strep7 | 0.810 | 0.645 | 1375 |
| Streptococcus pyogenes MGAS315 | 198466 | NCBI | Strep7 | 0.819 | 0.653 | 1527 |
| Streptococcus pyogenes MGAS8232 | 186103 | NCBI | Strep7 | 0.805 | 0.625 | 1485 |
| Streptococcus pyogenes SSI-1 | 193567 | NCBI | Strep7 | 0.814 | 0.643 | 1515 |
| Streptomyces avermitilis MA-4680 | 227882 | NCBI | Troph4 | 0.647 | 0.442 | 4903 |
| Streptomyces coelicolor A3(2) | 100226 | NCBI | Troph4 | 0.635 | 0.415 | 4772 |
| Sulfolobus solfataricus | 2287 | NCBI | Sulfo2 | 0.633 | 0.414 | 1885 |
| Sulfolobus tokodaii | 111955 | NCBI | Sulfo2 | 0.607 | 0.350 | 1714 |
| Synechococcus sp. WH 8102 | 84588 | JGI* | – | 0.642 | 0.438 | 1341 |
| Synechocystis sp. PCC 6803 | 1148 | NCBI | – | 0.661 | 0.485 | 2093 |
| Thermoanaerobacter tengcongensis | 119072 | NCBI | – | 0.845 | 0.629 | 2186 |
| Thermoplasma acidophilum | 2303 | NCBI | Therm2 | 0.633 | 0.392 | 938 |
| Thermoplasma volcanium | 50339 | NCBI | Therm2 | 0.638 | 0.422 | 957 |
| Thermosynechococcus elongatus BP-1 | 197221 | NCBI | – | 0.643 | 0.442 | 1591 |
| Thermotoga maritima | 2336 | NCBI | – | 0.800 | 0.745 | 1486 |
| Treponema pallidum | 160 | NCBI | – | 0.743 | 0.617 | 770 |
| Tropheryma whipplei str. Twist | 203267 | NCBI | Troph4 | 0.752 | 0.573 | 608 |
| Tropheryma whipplei TW08/27 | 218496 | NCBI | Troph4 | 0.762 | 0.589 | 597 |
| Ureaplasma urealyticum | 2130 | NCBI | – | 0.818 | 0.741 | 502 |
| Vibrio cholerae | 666 | NCBI | Wiggl19 | 0.670 | 0.485 | 2571 |
| Vibrio parahaemolyticus RIMD 2210633 | 223926 | NCBI | Wiggl19 | 0.666 | 0.478 | 3218 |
| Vibrio vulnificus CMCP6 | 216895 | NCBI | Wiggl19 | 0.678 | 0.505 | 3075 |
| Wigglesworthia glossinidia | 36870 | NCBI | Wiggl19 | 0.700 | 0.551 | 458 |
| Xanthomonas axonopodis | 190486 | NCBI | Xanth4 | 0.707 | 0.577 | 3048 |
| Xanthomonas campestris | 190485 | NCBI | Xanth4 | 0.707 | 0.577 | 2957 |
| Xylella fastidiosa 9a5c | 160492 | NCBI | Xanth4 | 0.683 | 0.515 | 1890 |
| Xylella fastidiosa Temecula1 | 183190 | NCBI | Xanth4 | 0.723 | 0.602 | 1470 |
| Yersinia pestis CO92 | 214092 | NCBI | Wiggl19 | 0.718 | 0.598 | 2789 |
| Yersinia pestis KIM | 187410 | NCBI | Wiggl19 | 0.667 | 0.486 | 2728 |

**Supplementary Table 4: Microarray experiments included in the analysis.** We only included experiments which measured levels of mRNA (excluding, for example, genomic hybridizations to compare strains). The Pearson correlation ($r$) for each pair of adjacent same-strand pairs was computed from normalized log-ratios. Because we only consider adjacent pairs, the results should not be sensitive to non-random layouts of the arrays.

For *E. coli K12*, *B. subtilis*, and *H. pylori*, we obtained data from the Stanford Microarray Database (http://genome-www.stanford.edu/microarray/), and we used only high-quality spots from each experiment (correlation between channels of at least 0.6, and a normalized log-ratio is present in the database). Microarray data for *H. pylori* was

for the Sydney strain 1 (SS1), whereas our database has sequence for the 26695 and J99 strains – for simplicity we compared the SS1 correlations to our predictions for strain 26695. For *C. trachomatis*, we used averages of normalized log-ratios from replicate experiments provided by Tracy Nicholson and Richard Stephens. Data for *Synechocystis* was obtained from KEGG (http://www.genome.ad.jp/kegg/expression/). Raw data from Suzuki et al (2001) was not available, so we used the normalized log-ratios provided. For the other *Synechocystis* experiments we normalized the data by splitting each array into 16 sectors and performing local regression of the log-ratio versus the sum of the logs ("M vs. A plots") within each sector. In all five genomes, we averaged multiple spots for a given gene (if available), subtracted out the mean log-ratio for each experiment before computing correlations, and required the pair of genes to be present in at least 10 arrays to report results. For *Halobacterium*, correlations for all adjacent pairs were provided by Richard Bonneau and Nitin Baliga.

For each experiment, the table shows: the source (database name and file name); our categorization of the experiments into conditions (used for jackknife statistics, see Fig. S9); the number of same-strand adjacent pairs where at least one gene changed (normalized $|log_2(ratio) \geq .3|$); for those pairs, a measure of agreement with our unsupervised predictions – the Spearman correlation between our $p$-value and the absolute difference between normalized log-ratios (more negative shows stronger agreement); and the publication describing the experiment.

**E.coli K12**

| Source | Condition | $n_{pairs}$ | $Spearman(p, \|\delta\|)$ | Publication |
|---|---|---|---|---|
| SMD:1278 | Control | 760 | -0.165 | Courcelle J et al.(2001) Genetics 158(1):41-64 |
| SMD:1285 | Control | 505 | -0.177 | Courcelle J et al.(2001) Genetics 158(1):41-64 |
| SMD:1290 | Control | 905 | -0.115 | Courcelle J et al.(2001) Genetics 158(1):41-64 |
| SMD:1292 | UV | 459 | -0.119 | Courcelle J et al.(2001) Genetics 158(1):41-64 |
| SMD:1592 | Control | 1415 | -0.267 | Khodursky AB et al.(2000) Proc Natl Acad Sci U S A 97:12170-5 |
| SMD:1593 | Control | 1190 | -0.254 | Khodursky AB et al.(2000) Proc Natl Acad Sci U S A 97:12170-5 |
| SMD:1595 | Control | 1387 | -0.264 | Khodursky AB et al.(2000) Proc Natl Acad Sci U S A 97:12170-5 |
| SMD:1597 | Control | 1296 | -0.256 | Khodursky AB et al.(2000) Proc Natl Acad Sci U S A 97:12170-5 |
| SMD:1637 | Tryptophan | 1300 | -0.279 | Khodursky AB et al.(2000) Proc Natl Acad Sci U S A 97:12170-5 |
| SMD:1638 | Tryptophan | 1881 | -0.258 | Khodursky AB et al.(2000) Proc Natl Acad Sci U S A 97:12170-5 |
| SMD:1639 | Tryptophan | 1288 | -0.251 | Khodursky AB et al.(2000) Proc Natl Acad Sci U S A 97:12170-5 |
| SMD:1641 | Tryptophan | 1172 | -0.206 | Khodursky AB et al.(2000) Proc Natl Acad Sci U S A 97:12170-5 |
| SMD:1642 | Tryptophan | 1295 | -0.197 | Khodursky AB et al.(2000) Proc Natl Acad Sci U S A 97:12170-5 |
| SMD:1643 | TrpMutant | 1519 | -0.254 | Khodursky AB et al.(2000) Proc Natl Acad Sci U S A 97:12170-5 |
| SMD:1644 | TrpMutant | 1953 | -0.240 | Khodursky AB et al.(2000) Proc Natl Acad Sci U S A 97:12170-5 |
| SMD:1646 | TrpMutant | 2009 | -0.269 | Khodursky AB et al.(2000) Proc Natl Acad Sci U S A 97:12170-5 |
| SMD:1647 | TrpMutant | 757 | -0.074 | Khodursky AB et al.(2000) Proc Natl Acad Sci U S A 97:12170-5 |
| SMD:1649 | TrpMutant | 1732 | -0.329 | Khodursky AB et al.(2000) Proc Natl Acad Sci U S A 97:12170-5 |
| SMD:1650 | TrpMutant | 1701 | -0.157 | Khodursky AB et al.(2000) Proc Natl Acad Sci U S A 97:12170-5 |
| SMD:1908 | Control | 500 | -0.170 | Courcelle J et al.(2001) Genetics 158(1):41-64 |
| SMD:1909 | Control | 601 | -0.204 | Courcelle J et al.(2001) Genetics 158(1):41-64 |
| SMD:1911 | lexA | 441 | -0.322 | Courcelle J et al.(2001) Genetics 158(1):41-64 |
| SMD:1912 | lexA | 438 | -0.223 | Courcelle J et al.(2001) Genetics 158(1):41-64 |
| SMD:1913 | lexA | 544 | -0.220 | Courcelle J et al.(2001) Genetics 158(1):41-64 |
| SMD:1914 | lexA | 406 | -0.229 | Courcelle J et al.(2001) Genetics 158(1):41-64 |
| SMD:1915 | lexA | 776 | -0.217 | Courcelle J et al.(2001) Genetics 158(1):41-64 |
| SMD:5265 | Control | 1529 | -0.200 | Khodursky AB et al.(2000) Proc Natl Acad Sci U S A 97:12170-5 |
| SMD:5266 | Tryptophan | 1850 | -0.227 | Khodursky AB et al.(2000) Proc Natl Acad Sci U S A 97:12170-5 |
| SMD:5268 | IndoleAcrylate | 1707 | -0.256 | Khodursky AB et al.(2000) Proc Natl Acad Sci U S A 97:12170-5 |
| SMD:5272 | IndoleAcrylate | 1706 | -0.236 | Khodursky AB et al.(2000) Proc Natl Acad Sci U S A 97:12170-5 |
| SMD:5273 | IndoleAcrylate | 1801 | -0.206 | Khodursky AB et al.(2000) Proc Natl Acad Sci U S A 97:12170-5 |
| SMD:5277 | IndoleAcrylate | 1792 | -0.183 | Khodursky AB et al.(2000) Proc Natl Acad Sci U S A 97:12170-5 |
| SMD:5278 | IndoleAcrylate | 1850 | -0.260 | Khodursky AB et al.(2000) Proc Natl Acad Sci U S A 97:12170-5 |
| SMD:5281 | IndoleAcrylate | 1956 | -0.214 | Khodursky AB et al.(2000) Proc Natl Acad Sci U S A 97:12170-5 |
| SMD:5284 | IndoleAcrylate | 2078 | -0.257 | Khodursky AB et al.(2000) Proc Natl Acad Sci U S A 97:12170-5 |
| SMD:5287 | IndoleAcrylate | 1993 | -0.243 | Khodursky AB et al.(2000) Proc Natl Acad Sci U S A 97:12170-5 |
| SMD:8377 | mRNAdecay | 1988 | -0.125 | Bernstein JA, et al. (2002) Proc Natl Acad Sci U S A 99(15):9697-702 |
| SMD:8378 | mRNAdecay | 1887 | -0.152 | Bernstein JA, et al. (2002) Proc Natl Acad Sci U S A 99(15):9697-702 |
| SMD:8379 | mRNAdecay | 1918 | -0.169 | Bernstein JA, et al. (2002) Proc Natl Acad Sci U S A 99(15):9697-702 |
| SMD:8380 | mRNAdecay | 1735 | -0.144 | Bernstein JA, et al. (2002) Proc Natl Acad Sci U S A 99(15):9697-702 |
| SMD:8536 | mRNAdecay | 1369 | -0.134 | Bernstein JA, et al. (2002) Proc Natl Acad Sci U S A 99(15):9697-702 |
| SMD:8538 | mRNAdecay | 1787 | -0.192 | Bernstein JA, et al. (2002) Proc Natl Acad Sci U S A 99(15):9697-702 |
| SMD:8540 | mRNAdecay | 1858 | -0.208 | Bernstein JA, et al. (2002) Proc Natl Acad Sci U S A 99(15):9697-702 |
| SMD:8542 | mRNAdecay | 2000 | -0.197 | Bernstein JA, et al. (2002) Proc Natl Acad Sci U S A 99(15):9697-702 |
| SMD:8582 | mRNAdecay | 1589 | -0.123 | Bernstein JA, et al. (2002) Proc Natl Acad Sci U S A 99(15):9697-702 |
| SMD:8584 | mRNAdecay | 1786 | -0.211 | Bernstein JA, et al. (2002) Proc Natl Acad Sci U S A 99(15):9697-702 |
| SMD:8586 | mRNAdecay | 2032 | -0.158 | Bernstein JA, et al. (2002) Proc Natl Acad Sci U S A 99(15):9697-702 |
| SMD:8589 | mRNAdecay | 2008 | -0.214 | Bernstein JA, et al. (2002) Proc Natl Acad Sci U S A 99(15):9697-702 |
| SMD:9395 | Control | 724 | -0.101 | Khodursky AB et al.(2000) Proc Natl Acad Sci U S A 97:12170-5 |

| Source | Condition | $n_{pairs}$ | $Spearman(p, \|\delta\|)$ | Publication |
|---|---|---|---|---|
| SMD:13838 | Control | 1752 | -0.280 | Lee K, et al. (2002) Mol Microbiol 43(6):1445-56 |
| SMD:13840 | rng | 1028 | -0.176 | Lee K, et al. (2002) Mol Microbiol 43(6):1445-56 |
| SMD:14829 | Leu | 906 | -0.169 | Tani TH, et al. (2002) Proc Natl Acad Sci U S A 99(21):13471-6 |
| SMD:14830 | lrp | 825 | -0.284 | Tani TH, et al. (2002) Proc Natl Acad Sci U S A 99(21):13471-6 |
| SMD:14831 | lrp | 1112 | -0.217 | Tani TH, et al. (2002) Proc Natl Acad Sci U S A 99(21):13471-6 |
| SMD:14832 | lrp | 671 | -0.152 | Tani TH, et al. (2002) Proc Natl Acad Sci U S A 99(21):13471-6 |
| SMD:15336 | Control | 1623 | -0.302 | Lee K, et al. (2002) Mol Microbiol 43(6):1445-56 |
| SMD:15337 | Control | 1221 | -0.276 | Lee K, et al. (2002) Mol Microbiol 43(6):1445-56 |
| SMD:15338 | Control | 1153 | -0.245 | Lee K, et al. (2002) Mol Microbiol 43(6):1445-56 |
| SMD:15339 | rne | 1655 | -0.355 | Lee K, et al. (2002) Mol Microbiol 43(6):1445-56 |
| SMD:15340 | rne | 1967 | -0.307 | Lee K, et al. (2002) Mol Microbiol 43(6):1445-56 |
| SMD:15341 | Control | 1846 | -0.203 | Lee K, et al. (2002) Mol Microbiol 43(6):1445-56 |
| SMD:15342 | Control | 842 | -0.332 | Lee K, et al. (2002) Mol Microbiol 43(6):1445-56 |
| SMD:15343 | rng | 802 | -0.209 | Lee K, et al. (2002) Mol Microbiol 43(6):1445-56 |
| SMD:25822 | Minimal | 1306 | -0.277 | Bernstein JA, et al. (2002) Proc Natl Acad Sci U S A 99(15):9697-702 |
| SMD:25825 | Minimal | 1446 | -0.264 | Bernstein JA, et al. (2002) Proc Natl Acad Sci U S A 99(15):9697-702 |
| SMD:25827 | Rich | 563 | -0.254 | Bernstein JA, et al. (2002) Proc Natl Acad Sci U S A 99(15):9697-702 |
| SMD:25831 | Rich | 530 | -0.280 | Bernstein JA, et al. (2002) Proc Natl Acad Sci U S A 99(15):9697-702 |
| SMD:32746 | rraA | 1033 | -0.134 | Lee K, et al.(2003)Cell 114:623-634 |
| SMD:32748 | rne | 1716 | -0.274 | Lee K, et al.(2003)Cell 114:623-634 |
| SMD:32749 | rne | 1934 | -0.282 | Lee K, et al.(2003)Cell 114:623-634 |
| SMD:32751 | rraA | 1495 | -0.335 | Lee K, et al.(2003)Cell 114:623-634 |
| SMD:32757 | rraA | 1591 | -0.117 | Lee K, et al.(2003)Cell 114:623-634 |
| SMD:32759 | rraA | 1233 | -0.160 | Lee K, et al.(2003)Cell 114:623-634 |
| SMD:32760 | rraA | 1384 | -0.136 | Lee K, et al.(2003)Cell 114:623-634 |

**B.subtilis**

| Source | Condition | $n_{pairs}$ | $Spearman(p, \|\delta\|)$ | Publication |
|---|---|---|---|---|
| SMD:22341 | Control | 971 | -0.325 | Helmann JD, et al.(2003) J Bacteriol 185(1):243-53 |
| SMD:22342 | Control | 710 | -0.332 | Helmann JD, et al.(2003) J Bacteriol 185(1):243-53 |
| SMD:22343 | Control | 912 | -0.264 | Helmann JD, et al.(2003) J Bacteriol 185(1):243-53 |
| SMD:22344 | Control | 1190 | -0.230 | Helmann JD, et al.(2003) J Bacteriol 185(1):243-53 |
| SMD:22345 | Control | 1153 | -0.230 | Helmann JD, et al.(2003) J Bacteriol 185(1):243-53 |
| SMD:22346 | Control | 1268 | -0.252 | Helmann JD, et al.(2003) J Bacteriol 185(1):243-53 |
| SMD:22347 | Control | 1178 | -0.262 | Helmann JD, et al.(2003) J Bacteriol 185(1):243-53 |
| SMD:22348 | Control | 1132 | -0.274 | Helmann JD, et al.(2003) J Bacteriol 185(1):243-53 |
| SMD:22349 | Control | 750 | -0.321 | Helmann JD, et al.(2003) J Bacteriol 185(1):243-53 |
| SMD:22350 | Control | 910 | -0.271 | Helmann JD, et al.(2003) J Bacteriol 185(1):243-53 |
| SMD:22351 | Control | 884 | -0.286 | Helmann JD, et al.(2003) J Bacteriol 185(1):243-53 |
| SMD:22352 | Control | 1381 | -0.338 | Helmann JD, et al.(2003) J Bacteriol 185(1):243-53 |
| SMD:22353 | Control | 1476 | -0.340 | Helmann JD, et al.(2003) J Bacteriol 185(1):243-53 |
| SMD:22354 | Control | 1549 | -0.343 | Helmann JD, et al.(2003) J Bacteriol 185(1):243-53 |
| SMD:22355 | Control | 157 | -0.219 | Helmann JD, et al.(2003) J Bacteriol 185(1):243-53 |
| SMD:22356 | Peroxide | 916 | -0.154 | Helmann JD, et al.(2003) J Bacteriol 185(1):243-53 |
| SMD:22357 | Peroxide | 1202 | -0.212 | Helmann JD, et al.(2003) J Bacteriol 185(1):243-53 |
| SMD:22358 | Peroxide | 1063 | -0.259 | Helmann JD, et al.(2003) J Bacteriol 185(1):243-53 |
| SMD:22359 | Peroxide | 1261 | -0.227 | Helmann JD, et al.(2003) J Bacteriol 185(1):243-53 |
| SMD:22360 | Peroxide | 966 | -0.278 | Helmann JD, et al.(2003) J Bacteriol 185(1):243-53 |
| SMD:22361 | Peroxide | 907 | -0.228 | Helmann JD, et al.(2003) J Bacteriol 185(1):243-53 |
| SMD:22362 | Peroxide | 633 | -0.127 | Helmann JD, et al.(2003) J Bacteriol 185(1):243-53 |
| SMD:22363 | Peroxide | 1220 | -0.280 | Helmann JD, et al.(2003) J Bacteriol 185(1):243-53 |
| SMD:22364 | Peroxide | 654 | -0.285 | Helmann JD, et al.(2003) J Bacteriol 185(1):243-53 |
| SMD:22365 | Control | 481 | -0.159 | Helmann JD, et al.(2003) J Bacteriol 185(1):243-53 |
| SMD:22366 | Peroxide | 965 | -0.278 | Helmann JD, et al.(2003) J Bacteriol 185(1):243-53 |
| SMD:22367 | Peroxide | 859 | -0.250 | Helmann JD, et al.(2003) J Bacteriol 185(1):243-53 |
| SMD:22368 | Peroxide | 951 | -0.282 | Helmann JD, et al.(2003) J Bacteriol 185(1):243-53 |
| SMD:22369 | Peroxide | 1118 | -0.267 | Helmann JD, et al.(2003) J Bacteriol 185(1):243-53 |
| SMD:22370 | Peroxide | 945 | -0.354 | Helmann JD, et al.(2003) J Bacteriol 185(1):243-53 |
| SMD:22371 | Peroxide | 1185 | -0.313 | Helmann JD, et al.(2003) J Bacteriol 185(1):243-53 |
| SMD:22372 | Peroxide | 628 | -0.251 | Helmann JD, et al.(2003) J Bacteriol 185(1):243-53 |
| SMD:22373 | Peroxide | 747 | -0.283 | Helmann JD, et al.(2003) J Bacteriol 185(1):243-53 |
| SMD:22374 | Peroxide | 890 | -0.294 | Helmann JD, et al.(2003) J Bacteriol 185(1):243-53 |
| SMD:22375 | Peroxide | 1614 | -0.297 | Helmann JD, et al.(2003) J Bacteriol 185(1):243-53 |
| SMD:22376 | Peroxide | 1609 | -0.300 | Helmann JD, et al.(2003) J Bacteriol 185(1):243-53 |
| SMD:22377 | Peroxide | 1109 | -0.352 | Helmann JD, et al.(2003) J Bacteriol 185(1):243-53 |
| SMD:22588 | Control | 276 | -0.341 | Helmann JD, et al.(2001) J Bacteriol 183(24):7318-28 |
| SMD:22589 | Control | 571 | -0.298 | Helmann JD, et al.(2001) J Bacteriol 183(24):7318-28 |
| SMD:22590 | Control | 521 | -0.218 | Helmann JD, et al.(2001) J Bacteriol 183(24):7318-28 |
| SMD:22591 | Control | 802 | -0.319 | Helmann JD, et al.(2001) J Bacteriol 183(24):7318-28 |
| SMD:22592 | Control | 542 | -0.292 | Helmann JD, et al.(2001) J Bacteriol 183(24):7318-28 |
| SMD:22593 | Control | 841 | -0.194 | Helmann JD, et al.(2001) J Bacteriol 183(24):7318-28 |
| SMD:22594 | Control | 437 | -0.105 | Helmann JD, et al.(2001) J Bacteriol 183(24):7318-28 |
| SMD:22595 | Control | 671 | -0.097 | Helmann JD, et al.(2001) J Bacteriol 183(24):7318-28 |
| SMD:22596 | Control | 1091 | -0.184 | Helmann JD, et al.(2001) J Bacteriol 183(24):7318-28 |
| SMD:22597 | Control | 371 | -0.156 | Helmann JD, et al.(2001) J Bacteriol 183(24):7318-28 |
| SMD:22598 | Heat | 1330 | -0.341 | Helmann JD, et al.(2001) J Bacteriol 183(24):7318-28 |
| SMD:22599 | Heat | 1755 | -0.322 | Helmann JD, et al.(2001) J Bacteriol 183(24):7318-28 |

| Source | Condition | $n_{pairs}$ | $Spearman(p, \|\delta\|)$ | Publication |
|---|---|---|---|---|
| SMD:22600 | Heat | 1555 | -0.320 | Helmann JD, et al.(2001) J Bacteriol 183(24):7318-28 |
| SMD:22601 | Heat | 1351 | -0.343 | Helmann JD, et al.(2001) J Bacteriol 183(24):7318-28 |
| SMD:22602 | Heat | 1574 | -0.340 | Helmann JD, et al.(2001) J Bacteriol 183(24):7318-28 |
| SMD:22604 | Heat | 1652 | -0.334 | Helmann JD, et al.(2001) J Bacteriol 183(24):7318-28 |
| SMD:22937 | Peroxide | 1275 | -0.275 | Helmann JD, et al.(2003) J Bacteriol 185(1):243-53 |
| SMD:22938 | Peroxide | 1317 | -0.262 | Helmann JD, et al.(2003) J Bacteriol 185(1):243-53 |
| SMD:22939 | Peroxide | 1229 | -0.271 | Helmann JD, et al.(2003) J Bacteriol 185(1):243-53 |
| SMD:22940 | Peroxide | 1393 | -0.328 | Helmann JD, et al.(2003) J Bacteriol 185(1):243-53 |
| SMD:22942 | Peroxide | 1373 | -0.428 | Helmann JD, et al.(2003) J Bacteriol 185(1):243-53 |
| SMD:22944 | Peroxide | 1198 | -0.409 | Helmann JD, et al.(2003) J Bacteriol 185(1):243-53 |
| SMD:22945 | Peroxide | 1269 | -0.337 | Helmann JD, et al.(2003) J Bacteriol 185(1):243-53 |
| SMD:22946 | Peroxide | 1206 | -0.351 | Helmann JD, et al.(2003) J Bacteriol 185(1):243-53 |
| SMD:22947 | Control | 743 | -0.332 | Helmann JD, et al.(2003) J Bacteriol 185(1):243-53 |
| SMD:22948 | Alcohol | 1258 | -0.281 | Helmann JD, et al.(2003) J Bacteriol 185(1):243-53 |
| SMD:22950 | Alcohol | 1162 | -0.311 | Helmann JD, et al.(2003) J Bacteriol 185(1):243-53 |
| SMD:22951 | Alcohol | 1049 | -0.282 | Helmann JD, et al.(2003) J Bacteriol 185(1):243-53 |
| SMD:22953 | Alcohol | 1225 | -0.354 | Helmann JD, et al.(2003) J Bacteriol 185(1):243-53 |
| SMD:22954 | Alcohol | 1318 | -0.330 | Helmann JD, et al.(2003) J Bacteriol 185(1):243-53 |
| SMD:22956 | Alcohol | 1393 | -0.369 | Helmann JD, et al.(2003) J Bacteriol 185(1):243-53 |
| SMD:22958 | Alcohol | 1401 | -0.318 | Helmann JD, et al.(2003) J Bacteriol 185(1):243-53 |
| SMD:22960 | Alcohol | 1345 | -0.315 | Helmann JD, et al.(2003) J Bacteriol 185(1):243-53 |
| SMD:22961 | Alcohol | 1298 | -0.337 | Helmann JD, et al.(2003) J Bacteriol 185(1):243-53 |
| SMD:22963 | Control | 1227 | -0.237 | Helmann JD, et al.(2003) J Bacteriol 185(1):243-53 |
| SMD:25863 | perR | 1559 | -0.274 | Helmann JD, et al.(2003) J Bacteriol 185(1):243-53 |
| SMD:25864 | perR | 1612 | -0.264 | Helmann JD, et al.(2003) J Bacteriol 185(1):243-53 |
| SMD:25865 | perR | 1566 | -0.301 | Helmann JD, et al.(2003) J Bacteriol 185(1):243-53 |
| SMD:25866 | perR | 1577 | -0.301 | Helmann JD, et al.(2003) J Bacteriol 185(1):243-53 |
| SMD:25867 | perR | 1543 | -0.335 | Helmann JD, et al.(2003) J Bacteriol 185(1):243-53 |
| SMD:25868 | perR | 1190 | -0.409 | Helmann JD, et al.(2003) J Bacteriol 185(1):243-53 |

**H.pylori 26695**

| Source | Condition | $n_{pairs}$ | $Spearman(p, \|\delta\|)$ | Publication |
|---|---|---|---|---|
| SMD:14840 | GrowthPhase | 811 | -0.165 | Thompson LJ, et al. (2003) Infect Immunity 71(5):2643-55 |
| SMD:14841 | GrowthPhase | 833 | -0.164 | Thompson LJ, et al. (2003) Infect Immunity 71(5):2643-55 |
| SMD:14842 | GrowthPhase | 795 | -0.190 | Thompson LJ, et al. (2003) Infect Immunity 71(5):2643-55 |
| SMD:14843 | GrowthPhase | 874 | -0.147 | Thompson LJ, et al. (2003) Infect Immunity 71(5):2643-55 |
| SMD:14844 | GrowthPhase | 820 | -0.113 | Thompson LJ, et al. (2003) Infect Immunity 71(5):2643-55 |
| SMD:14845 | GrowthPhase | 825 | -0.165 | Thompson LJ, et al. (2003) Infect Immunity 71(5):2643-55 |
| SMD:14847 | GrowthPhase | 793 | -0.086 | Thompson LJ, et al. (2003) Infect Immunity 71(5):2643-55 |
| SMD:14849 | GrowthPhase | 1039 | -0.056 | Thompson LJ, et al. (2003) Infect Immunity 71(5):2643-55 |
| SMD:14850 | GrowthPhase | 1006 | -0.113 | Thompson LJ, et al. (2003) Infect Immunity 71(5):2643-55 |
| SMD:15191 | GrowthPhase | 994 | -0.280 | Thompson LJ, et al. (2003) Infect Immunity 71(5):2643-55 |
| SMD:15192 | GrowthPhase | 914 | -0.290 | Thompson LJ, et al. (2003) Infect Immunity 71(5):2643-55 |
| SMD:15194 | GrowthPhase | 978 | -0.261 | Thompson LJ, et al. (2003) Infect Immunity 71(5):2643-55 |
| SMD:15195 | GrowthPhase | 1010 | -0.288 | Thompson LJ, et al. (2003) Infect Immunity 71(5):2643-55 |
| SMD:15196 | GrowthPhase | 995 | -0.247 | Thompson LJ, et al. (2003) Infect Immunity 71(5):2643-55 |
| SMD:15197 | GrowthPhase | 1030 | -0.198 | Thompson LJ, et al. (2003) Infect Immunity 71(5):2643-55 |
| SMD:15198 | GrowthPhase | 1081 | -0.218 | Thompson LJ, et al. (2003) Infect Immunity 71(5):2643-55 |
| SMD:15200 | GrowthPhase | 1109 | -0.202 | Thompson LJ, et al. (2003) Infect Immunity 71(5):2643-55 |
| SMD:21168 | pHlow | 759 | -0.182 | Merrell DS, et al(2003) Infect Immun 71(6):3529-39 |
| SMD:21169 | pHlow | 878 | -0.179 | Merrell DS, et al(2003) Infect Immun 71(6):3529-39 |
| SMD:21170 | pHlow | 866 | -0.210 | Merrell DS, et al(2003) Infect Immun 71(6):3529-39 |
| SMD:21171 | pHlow | 934 | -0.282 | Merrell DS, et al(2003) Infect Immun 71(6):3529-39 |
| SMD:21340 | pHlow | 697 | -0.112 | Merrell DS, et al(2003) Infect Immun 71(6):3529-39 |
| SMD:21341 | pHlow | 948 | -0.264 | Merrell DS, et al(2003) Infect Immun 71(6):3529-39 |
| SMD:21342 | pHlow | 934 | -0.268 | Merrell DS, et al(2003) Infect Immun 71(6):3529-39 |
| SMD:21343 | pHlow | 963 | -0.280 | Merrell DS, et al(2003) Infect Immun 71(6):3529-39 |
| SMD:31625 | GrowthPhase | 868 | -0.377 | Thompson LJ, et al. (2003) Infect Immunity 71(5):2643-55 |
| SMD:31626 | GrowthPhase | 891 | -0.347 | Thompson LJ, et al. (2003) Infect Immunity 71(5):2643-55 |
| SMD:31627 | GrowthPhase | 876 | -0.378 | Thompson LJ, et al. (2003) Infect Immunity 71(5):2643-55 |
| SMD:31628 | GrowthPhase | 867 | -0.373 | Thompson LJ, et al. (2003) Infect Immunity 71(5):2643-55 |
| SMD:31629 | GrowthPhase | 986 | -0.244 | Thompson LJ, et al. (2003) Infect Immunity 71(5):2643-55 |
| SMD:31630 | GrowthPhase | 1077 | -0.151 | Thompson LJ, et al. (2003) Infect Immunity 71(5):2643-55 |

**Synechocystis PCC 6803**

| Source | Condition | $n_{pairs}$ | $Spearman(p, \|\delta\|)$ | Publication |
|---|---|---|---|---|
| KEGG:ex0000022 | sycrp1 | 234 | -0.190 | Yoshimura H (2002) Mol Microbiol 43(4)843 |
| KEGG:ex0000023 | sycrp1 | 236 | -0.162 | Yoshimura H (2002) Mol Microbiol 43(4)843 |
| KEGG:ex0000024 | sycrp1 | 288 | -0.151 | Yoshimura H (2002) Mol Microbiol 43(4)843 |
| KEGG:ex0000030 | Redox | 428 | -0.195 | Hihara Y et al (2003) J Bacteriol 185(5):1719-25 |
| KEGG:ex0000038 | Redox | 353 | -0.244 | Hihara Y et al (2003) J Bacteriol 185(5):1719-25 |
| KEGG:ex0000044 | sycrp1 | 244 | -0.049 | Yoshimura H (2002) Mol Microbiol 43(4)843 |
| KEGG:ex0000049 | Redox | 752 | -0.134 | Hihara Y et al (2003) J Bacteriol 185(5):1719-25 |
| KEGG:ex0000050 | Redox | 704 | -0.161 | Hihara Y et al (2003) J Bacteriol 185(5):1719-25 |
| KEGG:ex0000051 | Redox | 706 | -0.177 | Hihara Y et al (2003) J Bacteriol 185(5):1719-25 |
| KEGG:ex0000052 | Redox | 751 | -0.233 | Hihara Y et al (2003) J Bacteriol 185(5):1719-25 |
| KEGG:ex0000140 | Light | 524 | -0.130 | Hihara Y et al 2001. Plant Cell, Vol. 13, 793-806 |
| KEGG:ex0000141 | Light | 763 | -0.158 | Hihara Y et al 2001. Plant Cell, Vol. 13, 793-806 |
| KEGG:ex0000142 | Light | 411 | -0.161 | Hihara Y et al 2001. Plant Cell, Vol. 13, 793-806 |
| KEGG:ex0000143 | Light | 422 | -0.236 | Hihara Y et al 2001. Plant Cell, Vol. 13, 793-806 |
| KEGG:ex0000144 | Light | 993 | -0.140 | Hihara Y et al 2001. Plant Cell, Vol. 13, 793-806 |
| KEGG:ex0000145 | Light | 950 | -0.092 | Hihara Y et al 2001. Plant Cell, Vol. 13, 793-806 |
| KEGG:ex0000146 | Light | 713 | -0.083 | Hihara Y et al 2001. Plant Cell, Vol. 13, 793-806 |
| KEGG:ex0000147 | Light | 624 | -0.151 | Hihara Y et al 2001. Plant Cell, Vol. 13, 793-806 |
| KEGG:ex0000148 | Light | 1005 | -0.077 | Hihara Y et al 2001. Plant Cell, Vol. 13, 793-806 |
| KEGG:ex0000149 | Light | 1026 | -0.130 | Hihara Y et al 2001. Plant Cell, Vol. 13, 793-806 |
| KEGG:ex0000152 | Light | 702 | -0.189 | Hihara Y et al 2001. Plant Cell, Vol. 13, 793-806 |
| KEGG:ex0000153 | Light | 840 | -0.165 | Hihara Y et al 2001. Plant Cell, Vol. 13, 793-806 |
| KEGG:ex0000154 | Redox | 730 | -0.185 | Hihara Y et al (2003) J Bacteriol 185(5):1719-25 |
| KEGG:ex0000155 | Redox | 550 | -0.136 | Hihara Y et al (2003) J Bacteriol 185(5):1719-25 |
| KEGG:ex0000156 | Redox | 608 | -0.112 | Hihara Y et al (2003) J Bacteriol 185(5):1719-25 |
| KEGG:ex0000157 | Redox | 547 | -0.177 | Hihara Y et al (2003) J Bacteriol 185(5):1719-25 |
| KEGG:ex0000158 | Redox | 333 | -0.352 | Hihara Y et al (2003) J Bacteriol 185(5):1719-25 |
| KEGG:ex0000159 | Redox | 320 | -0.349 | Hihara Y et al (2003) J Bacteriol 185(5):1719-25 |
| KEGG:ex0000160 | Light | 321 | -0.169 | Hihara Y et al 2001. Plant Cell, Vol. 13, 793-806 |
| KEGG:ex0000161 | Light | 367 | -0.189 | Hihara Y et al 2001. Plant Cell, Vol. 13, 793-806 |
| KEGG:ex0000162 | Light | 622 | -0.193 | Hihara Y et al 2001. Plant Cell, Vol. 13, 793-806 |
| KEGG:ex0000163 | Light | 640 | -0.164 | Hihara Y et al 2001. Plant Cell, Vol. 13, 793-806 |
| KEGG:ex0000164 | Light | 665 | -0.218 | Hihara Y et al 2001. Plant Cell, Vol. 13, 793-806 |
| KEGG:ex0000165 | Light | 537 | -0.135 | Hihara Y et al 2001. Plant Cell, Vol. 13, 793-806 |
| KEGG:ex0000166 | Redox | 1168 | -0.160 | Hihara Y et al (2003) J Bacteriol 185(5):1719-25 |
| KEGG:ex0000167 | Redox | 1185 | -0.145 | Hihara Y et al (2003) J Bacteriol 185(5):1719-25 |
| KEGG:ex0000168 | Light | 662 | -0.170 | Hihara Y et al 2001. Plant Cell, Vol. 13, 793-806 |
| KEGG:ex0000169 | Light | 1039 | -0.132 | Hihara Y et al 2001. Plant Cell, Vol. 13, 793-806 |
| KEGG:ex0000832 | Redox | 799 | -0.253 | Hihara Y et al (2003) J Bacteriol 185(5):1719-25 |
| KEGG:ex0000833 | Redox | 869 | -0.239 | Hihara Y et al (2003) J Bacteriol 185(5):1719-25 |
| KEGG:ex0000834 | Redox | 1196 | -0.188 | Hihara Y et al (2003) J Bacteriol 185(5):1719-25 |
| KEGG:ex0000835 | Redox | 1228 | -0.197 | Hihara Y et al (2003) J Bacteriol 185(5):1719-25 |
| KEGG:ex0000836 | Redox | 853 | -0.255 | Hihara Y et al (2003) J Bacteriol 185(5):1719-25 |
| KEGG:ex0000837 | Redox | 959 | -0.252 | Hihara Y et al (2003) J Bacteriol 185(5):1719-25 |
| KEGG:ex0000838 | Redox | 1063 | -0.247 | Hihara Y et al (2003) J Bacteriol 185(5):1719-25 |
| KEGG:ex0000839 | Redox | 1276 | -0.217 | Hihara Y et al (2003) J Bacteriol 185(5):1719-25 |
| KEGG:exn0000001 | Cold | 1365 | -0.062 | Suzuki I et al 2001. Mol. Microbiology 401(1):235-244 |
| KEGG:exn0000002 | hik33 | 1013 | -0.075 | Suzuki I et al 2001. Mol. Microbiology 401(1):235-244 |
| KEGG:exn0000003 | Cold | 1090 | -0.027 | Suzuki I et al 2001. Mol. Microbiology 401(1):235-244 |

**C.trachomatis**

| Source | Condition | $n_{pairs}$ | $Spearman(p, \|\delta\|)$ | Publication |
|---|---|---|---|---|
| TN & RS | IFNgamma2 | 210 | -0.044 | – |
| TN & RS | AlphaKetoGlutamate | 490 | -0.189 | – |
| TN & RS | Glutamate | 227 | 0.016 | – |
| TN & RS | Heat | 288 | -0.278 | – |
| TN & RS | Heat | 158 | -0.238 | – |
| TN & RS | Iron | 460 | -0.156 | – |
| TN & RS | Penicillin | 330 | -0.359 | – |
| TN & RS | Growth | 551 | -0.191 | Nicholson TL et al 2003 J Bact 185(10):3179-89 |
| TN & RS | Growth | 525 | -0.253 | Nicholson TL et al 2003 J Bact 185(10):3179-89 |
| TN & RS | Growth | 374 | -0.176 | Nicholson TL et al 2003 J Bact 185(10):3179-89 |
| TN & RS | Growth | 566 | -0.251 | Nicholson TL et al 2003 J Bact 185(10):3179-89 |
| TN & RS | Tryptophan | 307 | -0.036 | – |

**Halobacterium NRC-1**

| Source | Condition | $n_{pairs}$ | $Spearman(p, \lvert\delta\rvert)$ | Publication |
|---|---|---|---|---|
| RN & NB | UV resistance (5 experiments) | – | – | – |
| RN & NB | Light/dark response (3 experiments) | – | – | – |
| RN & NB | bop/bat genetic perturbations (3 experiments) | – | – | Baliga NS et al. (2002) PNAS 99:14913-8 |
| RN & NB | day/night entrainment (33 experiments) | – | – | – |