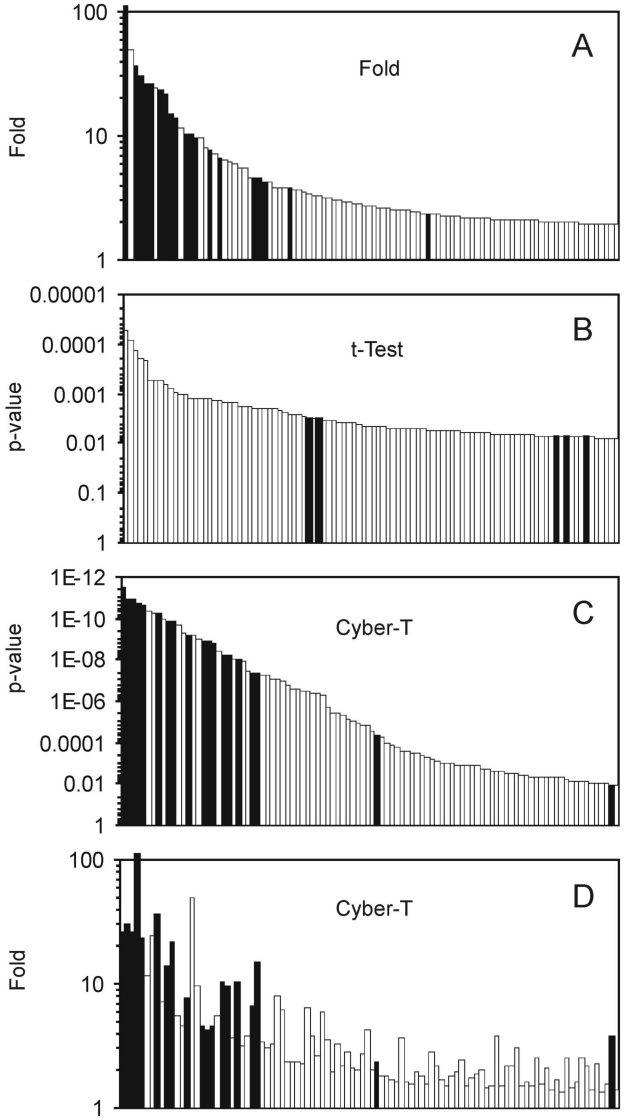
SUPPLEMENTARY MATERIAL

COMPARISON OF METHODS FOR THE DEFINITION OF THE COMK-REGULON

In the analysis of the macroarrays the question arises how the ComK-regulon is best defined, *i.e.* which genes can be considered to be significantly activated by ComK and which genes are not influenced by this regulator. A comparison of the transcription profiles of the *comK* knockout strain (BV2004) and the control strain (BV2012) revealed 94 genes with a greater than 2-fold difference in expression level. As shown in the first bar diagram of Fig. 1, almost all of the known ComK-activated genes (summarized in Table 1) were up-regulated more than 2-fold. Three genes known to be ComK regulated were missing from this list: *addA, addB*, and *uvrB* did not meet the criterium. It is noteworthy that these genes are also missing in two other recently published papers on the ComK-regulon (1;2). Generally, a 2-fold difference is assumed to be significant in DNA-array experiments. Arfin et al., however, convincingly showed that the criterion of a 2-fold difference in expression levels is disputable (3). They used a statistical approach to discriminate between regulated and non-regulated genes and showed that a 1.5-fold difference in expression can be significant, whereas in several cases differences of more than 2-fold appeared to be unreliable. A major difficulty with applying statistical analyses on array data is the small number of replicates, which results in poor estimates of variance and a correspondingly poor performance of a *t* test, as illustrated in Fig. 2 B. Long et al. developed a statistical program, Cyber-T, incorporating a Bayesian prior to improve the *t* test for DNA-array measurements (4). They assumed that genes with similar expression levels have similar measurement errors, and that a set of such genes can be used as pseudo-replicates to increase the reliability of the variance (for a comprehensive discussion see (5)). As shown in Fig. 1 C, when Cyber-T was used to analyze the data set, the majority of known ComK-activated genes clustered at the high-reliability end of the bar diagram, comparable to the clustering seen when genes were sorted on fold differences in expression (Fig. 1 A). The last bar diagram of Fig. 1 gives an indication of the correlation between expression differences and p-values measured by Cyber-T. Cyber-T thus provides us with a statistical method to select ComK activated genes, which performs at least as well as a selection based on expression differences.

Figure S1. Genes differently expressed between wild type and *comK* mutant strain, and sorted on fold differences in expression, or sorted on significance of differences in expression (p-values). Bar diagrams display the first 100 genes after sorting. Closed bars indicate known ComK-activated genes. A) Genes sorted on fold differences in expression, B) genes sorted on p-values calculated by a Student *t* test, C) genes sorted on p-values calculated by Cyber-T, D) sorted as C, but displaying the fold differences in expression.



Reference List

1. Berka,R.M., Hahn,J., Albano,M., Draskovic,I., Persuh,M., Cui,X., Sloma,A., Widner,W. and Dubnau,D. (2002) Microarray analysis of the Bacillus subtilis K-state: genome-wide expression changes dependent on ComK. *Mol.Microbiol.*, **43**, 1331-1345.

2. Ogura,M., Yamaguchi,H., Kobayashi,K., Ogasawara,N. , Fujita,Y. and Tanaka,T. (2002) Whole-Genome Analysis of Genes Regulated by the Bacillus subtilis Competence Transcription Factor ComK. *J.Bacteriol.*, **184**, 2344-2351.

3. Arfin,S.M., Long,A.D., Ito,E.T., Tolleri,L., Riehle,M.M., Paegle,E.S. and Hatfield,G.W. (2000) Global gene expression profiling in Escherichia coli K12. The effects of integration host factor. *J.Biol.Chem.*, **275**, 29672-29684.

4. Long,A.D., Mangalam,H.J., Chan,B.Y., Tolleri,L., Hatfield,G.W. and Baldi,P. (2001) Improved statistical inference from DNA microarray data using analysis of variance and a Bayesian statistical framework. Analysis of global gene expression in Escherichia coli K12. *J.Biol.Chem.*, **276**, 19937-19944.

5. Baldi,P. and Long,A.D. (2001) A Bayesian framework for the analysis of microarray expression data: regularized t -test and statistical inferences of gene changes. *Bioinformatics.*, **17**, 509-519.

EFFECT OF AN ANTIBIOTIC RESISTANCE MARKER ON TRANSCRIPTION PROFILES

Experiments were carried out with RNA isolated form wild type cells and from cells mutated in *comK*. For convenience, comK was mutated by insertion of a spectinomycin resistance marker. So in fact this strain contains two genetic modifications; a disrupted comK gene and the presence of an antibiotic resistance gene. In order to detect genuine ComK-dependent expression, we examined to which extend the presence of a spectinomycin resistance marker alters the global *B. subtilis* transcription profile. A *B. subtilis* strain was constructed in which the spectinomycin marker was inserted at a suitable place in the genome, such that the expression of no gene nor operon was affected. We chose the *pks* locus for integration. The *pks* operon is one of the largest operons in *B. subtilis* and encodes components of a polyketide synthetase implicated in synthesis of the antibiotic difficidin (1). Mutations in this locus showed no phenotype and so far production of difficidin by *B. subtilis* strain 168 has not been reported (2). The spectinomycin marker was integrated between the 3'-end of the last gene (*pksS*) and the terminator of the operon, and had the same transcription direction. The adjacent *pksR* gene is transcribed in the opposite direction.

Both the wild-type strain (8G5) and the strain containing the spectinomycin marker at the pks locus (BV2012) were grown in competence medium (without spectinomycin) according to a two-step competence protocol and cells were harvested when the competence stage was reached. The experimental set-up was the same as for the determination of the ComK-regulon (see Materials and Methods of the original article for details). 16 genes showed more than 2 fold difference in expression due to the presence of a spectinomycin marker. We used Cyber-T to calculate the significance of our array-data and Table 1 presents a summary of genes with > 1.5 fold difference in expression, due to the presence of a spectinomycin marker, and p-value < 0.05 (see for an extensive discussion on folds versus p-values Long et al. (3)). Only 15 genes met our selection criteria of which 3 genes showed a difference in expression of more than 2 fold. The lower expression of *pksR* is likely to be related to read through from the adjacent spectinomycin resistance gene. Clearly, the presence of the spectinomycin resistance gene has only a marginal effect on the global transcription pattern, and it is unlikely that this marker will obscure the global transcription effect of a mutated *comK*. However, since at least some effects were observed, we decided to use strain BV2012 as reference "wild-type" strain in the global transcription profiling of a *comK* mutant.

Table S1. Effect of a spectinomycin resistance marker (Spc) on global gene expression in *B. subtilis*. The transcription profile of a wild type strain (wt) was compared to the transcription profiles of (i) strain BV2012 containing a Spc marker at the *pks* locus (Spc), and (ii) strain BV2004 containing a Spc marker inserted into *comK* (comK-Spc). Only genes with p-values < 0.016 are displayed (see main text for details). p-Values were determined using Cyber-T. Positive fold differences indicate increased gene expression due to the presence of the Spc marker. Negative fold differences indicate decreased gene expression due to the presence of the Spc marker.

Gene	Fold Spc / wt	Fold comK-Spc / wt	Description
carA	3.42	5.02	Carbamoyl-phosphate transferase-arginine (subunit A)
tnrA	2.07	1.85	Transcriptional pleiotropic regulator involved in nitrogen regulation
ybbM	1.88	1.97	Similar to hypothetical proteins
yclP	1.64	1.78	Similar to ferrichrome ABC transporter (ATP-binding protein)
yfmT	1.54	1.73	Similar to benzaldehyde dehydrogenase
yczI	1.52	1.63	No similarity to other proteins
pksR	-1.91	-	Polyketide synthase

Reference List

1. Kunst,F., Ogasawara,N., Moszer,I., Albertini,A.M., Alloni,G., Azevedo,V., Bertero,M.G., Bessieres,P., Bolotin,A., Borchert,S. *et al.* (1997) The complete genome sequence of the gram-positive bacterium *Bacillus subtilis*. *Nature*, **390**, 249-256.

2. Scotti,C., Piatti,M., Cuzzoni,A., Perani,P., Tognoni,A., Grandi,G., Galizzi,A. and Albertini,A.M. (1993) A Bacillus subtilis large ORF coding for a polypeptide highly similar to polyketide synthases. *Gene*, **130**, 65-71.

3. Long,A.D., Mangalam,H.J., Chan,B.Y., Tolleri,L., Hatfield,G.W. and Baldi,P. (2001) Improved statistical inference from DNA microarray data using analysis of variance and a Bayesian statistical framework. Analysis of global gene expression in Escherichia coli K12. *J.Biol.Chem.*, **276**, 19937-19944.

IDENTIFICATION OF REPRESSED GENES

Both the wild-type strain (8G5) and the strain containing the spectinomycin marker at the pks locus (BV2012) were grown in competence medium (without spectinomycin) according to a two-step competence protocol and cells were harvested when the competence stage was reached. The experimental set-up was the same as for the determination of the ComK-regulon (see Materials and Methods of the original article for details).
In our study we found 6 genes (Table 1), which were repressed significantly when the *flgM* standards were applied after analysis with Cyber-T (1). The repression of one of those genes, *pksR*, is presumably a polar effect of the insertion of the Spc marker adjacent to *pksR* in the control strain BV2012. The differences in expression of the other five genes are low. Whether these marginal differences originate from the competent or non-competent cell fraction, or whether they are relevant for competence development, is unknown. According to experiments done in the BSFA consortium, at least one of the genes (*ylqD*) is not required for competence.

Table S2. Genes repressed by ComK. The expression of the selected genes showed at least 1.4-fold differences with p-values < 0.017. Genes are sorted according to expression differences. The functional descriptions are based on the Subtilist Web Server (http://genolist.pasteur.fr/SubtiList/) information. The distance to the first upstream K-box is given in base pairs, together with type and bp match of the K-box.

Gene	Description	Fold	Involved in competence	Distance to K-box	K-box type
pksR	Polyketide synthase	-1.7			
ylqD	Similar to hypothetical proteins	-1.6	no (1)	622	II-13
ykaA	No similarity to other proteins	-1.5		169	II-13
thiF	Thiamine biosynthesis	-1.4		2769	II-13
rapF	Response regulator aspartate phosphatase	-1.4		9733	III-13
prs	Phosphoribosyl pyrophosphate synthetase	-1.4		9021	II-13

(1) BSFA, *Bacillus subtilis* functional analyses programme. Transformation percentages < 10% were considered as disturbed competence.

Reference List

1. Long,A.D., Mangalam,H.J., Chan,B.Y., Tolleri,L., Hatfield,G.W. and Baldi,P. (2001) Improved statistical inference from DNA microarray data using analysis of variance and a Bayesian statistical framework. Analysis of global gene expression in Escherichia coli K12. *J.Biol.Chem.*, **276**, 19937-19944.