Defining the structure of the general stress regulon of *Bacillus subtilis* using targeted microarray analysis and random forest classification

Priyanka Nannapaneni,1 Falk Hertwig,1† Maren Depke,1 Michael Hecker,2 Ulrike Mäder,1 Uwe Völker,1 Leif Steil1 and Sacha A. F. T. van Hijum1,3,4

1Interfakultäres Institut für Genetik und Funktionelle Genomforschung, Ernst-Moritz-Arndt-Universität Greifswald, Germany
2Institut für Mikrobiologie, Ernst-Moritz-Arndt-Universität Greifswald, Germany
3Radboud University Nijmegen Medical Centre, Center for Molecular and Biomolecular Informatics, Nijmegen Center for Molecular Life Sciences, PO Box 9101, 6500 HB Nijmegen, The Netherlands
4NIZO Food Research, PO Box 20, 6710 BA Ede, The Netherlands

The structure of the SigB-dependent general stress regulon of *Bacillus subtilis* has previously been characterized by proteomics approaches as well as DNA array-based expression studies. However, comparing the SigB targets published in three previous major transcriptional profiling studies it is obvious that although each of them identified well above 100 target genes, only 67 were identified in all three studies. These substantial differences can likely be attributed to the different strains, growth conditions, microarray platforms and experimental setups used in the studies. In order to gain a better understanding of the structure of this important regulon, a targeted DNA microarray analysis covering most of the known SigB-inducing conditions was performed, and the changes in expression kinetics of 252 potential members of the SigB regulon and appropriate control genes were recorded. Transcriptional data for the *B. subtilis* wild-type strain 168 and its isogenic *sigB* mutant BSM29 were analysed using random forest, a machine learning algorithm, by incorporating the knowledge from previous studies. This analysis revealed a strictly SigB-dependent expression pattern for 166 genes following ethanol, butanol, osmotic and oxidative stress, low-temperature growth and heat shock, as well as limitation of oxygen or glucose. Kinetic analysis of the data for the wild-type strain identified 30 additional members of the SigB regulon, which were also subject to control by additional transcriptional regulators, thus displaying atypical SigB-independent induction patterns in the mutant strain under some of the conditions tested. For 19 of these 30 SigB regulon members, published reports support control by secondary regulators along with SigB. Thus, this microarray-based study assigns a total of 196 genes to the SigB-dependent general stress regulon of *B. subtilis*.

INTRODUCTION

As an organism living in the upper layer of soil, *Bacillus subtilis* is exposed to a wide range of transitory stress and starvation conditions, and many of these activate the alternative sigma factor SigB. The collection of SigB-activating stress conditions includes sudden environmental stresses such as heat, acid, ethanol, butanol and osmotic stress, Mn^{2+}, nitric oxide (NO), sodium nitroprusside (SNP) and blue light; energy stresses such as starvation for glucose, phosphate and oxygen; inhibitors that collectively trigger a decrease in the ATP pool [azide, carbonyl cyanide m-chlorophenylhydrazone (CCCP), mycophenolic acid], addition of antibiotics such as vancomycin and bacitracin; as well as growth at extreme temperatures, both high and low (Hecker et al., 2007; Price, 2000, 2002, 2011). The increased expression of the SigB regulon provides *B. subtilis* cells with a non-specific, multiple stress resistance against a wide range of stresses (Gaidenko & Price, 1998; Höper et al., 2005; Völker et al., 1999). SigB-dependent gene

1Present address: Lund Strategic Research Center for Stem Cell Biology, Department of Laboratory Medicine, Lund University, Sweden.

Abbreviations: ANOVA, analysis of variance; OOB, out-of-bag; RF, random forest.

The complete microarray dataset discussed in this paper is available from the NCBI Gene Expression Omnibus (GEO) under accession number GSE32895.

Three supplementary figures and six supplementary tables are available with the online version of this paper.
expression also contributes to growth and survival at low temperatures and during extended stationary phases (Brigulla et al., 2003; Méndez et al., 2004). Though there are several studies supporting the important role of this general stress regulon, a detailed study of the functions of individual regulon members is required to comprehensively understand the role of the regulon itself in cellular stress management.

A comprehensive comparative phenotypic screening analysis of a B. subtilis wild-type strain, a sigB mutant and 94 individual mutants of candidate SigB-dependent genes revealed that 85% of the mutants carrying knockouts in single SigB-dependent genes displayed increased stress sensitivity during severe ethanol stress, heat shock and/or salt stress, and/or impaired growth at low temperature (Höper et al., 2005). However, a more comprehensive screening of the function of SigB-dependent genes requires precise knowledge of the structure of the SigB regulon.

The B. subtilis general stress response has been extensively studied since the 1990s. Several research groups have identified regulon members by various approaches, including transposon mutagenesis (Boylan et al., 1991, 1992, 1993), gel-based proteome analysis or consensus promoter search-driven extensive oligonucleotide hybridization screening (Petersohn et al., 1999a; Völker et al., 1994). After the genome sequence of B. subtilis became available in 1997 (Kunst et al., 1997), a comprehensive search for SigB regulon members was performed in three genome-wide transcriptional profiling studies (Helmann et al., 2001; Petersohn et al., 2001; Price et al., 2001). These three DNA array studies were performed to identify SigB regulon members by testing subsets of well-described stress conditions. Price et al. (2001) studied ethanol stress (5%, v/v) using a custom macroarray with PCR products spotted on nylon membranes. They compared a wild-type strain and a sigB mutant strain cultivated in complex Luria–Bertani (LB) medium and postulated 193 genes to be SigB-dependent. Helmann et al. (2001) conducted experiments exposing a wild-type strain to heat shock (shift from 37 to 48 °C) in minimal medium and used DNA microarray glass slides. That study identified 123 genes as members of the SigB regulon. Petersohn et al. (2001) compared a wild-type strain and its isogenic sigB mutant after ethanol stress (4%, w/v), heat shock (shift from 37 to 48 °C) and osmotic stress (4 to 10%, v/v, NaCl). They assayed transcriptional patterns using DNA macroarrays containing PCR products of the whole genome and 124 genes were claimed to be SigB-dependent. When comparing the different sets of SigB regulon members postulated by these three studies, it became obvious that the overlap was rather small. Out of the total of 261 genes postulated in these three studies, only 67 genes were described as being SigB-dependent in all three studies (Fig. 1). Potential reasons for the greatly varying sets of genes identified are the different array platforms, strains, media and experimental conditions tested. Therefore, the structure of the SigB-dependent general stress regulon of B. subtilis remains to be elucidated. Here we present the SigB targets derived from a comprehensive targeted comparative microarray study of the expression changes observed in the B. subtilis wild-type strain 168 and its isogenic sigB mutant BSM29 using a wide range of SigB-inducing conditions. The expression patterns and the list of 196 SigB-dependent genes provide a resource for further testing of the individual contributions of these genes to the development of stress resistance in B. subtilis.

**METHODS**

**Strains and culture conditions.** In this work, the following B. subtilis strains were used: the wild-type strain 168 Marburg (trpC2) (Burkholder & Giles, 1947) and its isogenic mutant BSM29 (trpC2 sigB::spec), which was constructed by insertion of a spectinomycin-resistance cassette into sigB (Brigulla et al., 2003). The strains were precultured in LB medium, in the case of BSM29 supplemented with 50 μg spectinomycin ml⁻¹. For the stress experiments, a defined minimal medium was selected (Stülke et al., 1993).

Expression data were generated in a time-series experiment, querying the wild-type and the sigB mutant. Stress conditions tested were: environmental stress (ethanol stress, butanol stress, heat stress, osmotic stress); energy stress (O₂ limitation and glucose limitation); oxidative stress and low-temperature growth. Sampling started in the exponential growth phase, at OD₅₄₀ 0.45–0.55 for both strains (corresponding to time 0). Various stresses were applied to the cultures incubated at 37 °C with shaking at 210 r.p.m.: ethanol stress [the culture was treated with a final concentration of 4% (v/v) ethanol (Völker et al., 1995)]; salt stress [solid sodium chloride was added to a final concentration of 4% (w/v) (Petersohn et al., 2001)]; butanol stress [the culture was treated with a final concentration of 1% (v/v) butanol (Helmann et al., 2003)]; heat stress [the culture was shifted from 37 to 48 °C (Helmann et al., 2001)]; and oxidative stress [the culture was treated with hydrogen peroxide at a final sublethal concentration of 60 μM (Helmann et al., 2003)]. Samples were taken...
at regular time intervals of 5, 10, 15 and 20 min after the imposition of the environmental stress factors. For low-temperature growth, the cultures were grown in minimal medium at 37 °C to OD540 0.1 and then the cultures were shifted to 15 °C (Brigulla et al., 2003). Afterwards, the cultures were sampled at OD540 0.9 and 1.0. Glucose limitation was triggered by growing the strains with growth-limiting amounts of 0.05 % (w/v) glucose at 37 °C (Völker et al., 1995).

Oxygen limitation was induced by reducing the shaking speed at OD540 0.3 from 210 to 50 r.p.m. The sampling for both glucose and oxygen limitation was conducted at the time intervals of 15, 30, 45, 60 and 90 min after the induction of the limitation. Each cultivation was performed in duplicate, except for butanol stress, where only one cultivation was performed. Total RNA was isolated according to the method described by Petersohn et al. (2001).

**Array design.** Of the 350 genes included in the array probe design, 252 genes had been described as SigB-dependent by one or more of the studies of Price, Petersohn and Helmann (Helmann et al., 2001; Petersohn et al., 2001; Price et al., 2001). Just nine genes from the study of Petersohn et al. (2001) were missing from the array for technical reasons. The array also included 52 genes predicted to be SigB-dependent by other published studies (Petersohn et al., 1999b, 2001; Price et al., 2001), 10 additional genes of interest based on unpublished data, and 36 control genes not known to be controlled by SigB (Supplementary Table S1). ORF sequences for the 350 selected genes were exported from GenBank and the probe design was provided by GENESCAN Europe (Freiburg, Germany). Isothermal probes of 50 nt in length were derived from the ORF sequences and spotted in triplicate. Probes were covalently attached to a hydrophobic substrate on the surface of glass slides, as described elsewhere (Waldmüller et al., 2002). cDNA was synthesized and labelled as described previously (Jürgen et al., 2005).

**Array hybridization and spot quantification.** Both wild-type and mutant RNA samples were labelled with Cy3 and Cy5 dyes, depending on the hybridization scheme (Supplementary Fig. S1). The Cy3- and Cy5-labelled cDNA solutions (60 μl) were then denatured for 1 min at 100 °C, cooled down for 5 min on ice and immediately pipetted onto the array. Two gene frames (ABgene) had been glued previously on top of the hybridization region, creating a 240 μl chamber on the microarray according to the manufacturer’s instructions. The resulting chamber was filled with hybridization solution and closed with a flexible coverslip (ABgene) without leaving any air bubbles. The microarray was then inserted into a 50 ml tube covered with aluminium foil and incubated under rotation at 60 °C for 16 h. After hybridization, the gene frames with coverslips were carefully removed and the array was reinserted into the 50 ml tube. Subsequently, the microarray in the 50 ml tubes was subjected to washing in the following buffers: 2 × SSC, 0.1 % SDS for 5 min at 65 °C, 0.5 × SSC for 5 min at 25 °C and 0.1 × SSC for 10 min at 25 °C. Finally, the arrays were placed in an empty 50 ml tube and centrifuged at 650 g for drying. Slides were then scanned using a ScanArray Express microarray scanner (Perkin Elmer Lifesciences; Cy3 at 543 nm, Cy5 at 633 nm). The fluorescence was measured by a photomultiplier in both channels and a digitized image was stored for each. From these images, the signal intensities of individual spots were calculated with the Array Vision 6.0 program (GE Healthcare). The software extracted the median trimmed density after subtraction of local background. Thereafter, the three replicate spots were flagged as present (P), marginal (M) or absent (A). A spot was flagged as present (P) if the S : N ratio was larger than two and the difference of the median spot intensities between replicate spots was less than 1.25-fold. Marginal spots had to meet the same S : N ratio, but the difference of the median spot intensities between replicate spots had to be less than 1.5-fold. Spots with an S : N ratio of less than two or a difference of the median spot intensities between replicate spots of greater than 1.5-fold were flagged as absent. Replicate spot intensities were combined using the median of the spots flagged as present or marginal. If none of the replicate spots was flagged as present or marginal, the spot intensity values of the absent flagged spots were taken to obtain a complete data matrix for the random forest (RF) algorithm. Three replicates for each tested condition were analysed; among them, one was the dye swap and the other two were biological replicates.

**Microarray data processing.** The microarray hybridization design consisted of different stress conditions and time points for both the wild-type and the sigB knockout strain. Correction for systematic effects (e.g. dye effects) in dual-dye microarray data are usually achieved using Lowess normalization. In the current hybridization design, the wild-type and sigB knockout were always cohybridized; consequently, the gene expression could vary widely depending on the stress condition and time point. The assumption of Lowess normalization that a fraction of genes/probes (50 % in our study) has unchanged expression can therefore be challenged for some of the hybridizations performed here. To this end, we created so-called synthetic slides, containing conditions that are expected to have similar gene expression patterns, or conditions that are equivalent (over 50% of the probes should have similar signals after normalization; Supplementary Fig. S2). We determined the equivalence of samples using correlation analysis. Across stress time points, we treated the signals of each channel (either Cy3 or Cy5) as separate datasets. The gene expression intensities of the 350 spotted genes for the wild-type and sigB knockout strain datasets were correlated using Pearson’s product moment. We then paired datasets into synthetic slides, provided that their Pearson’s correlation was above 0.8. To normalize for systematic effects the probe signals were normalized by Lowess (f=0.5) using MicroPrep (van Hjium et al., 2003). Next, ratios of gene expression for each gene of the wild-type strain were calculated with respect to the sigB knockout strain. These ratios and normalized gene expression values were used for further analysis.

**RF classification.** The RF classifier package (version 4.5-35) (Breiman, 2001) implemented in the statistical programming language R (version 2.8.0) was used to train RF classifiers for the separation of SigB from non-SigB members based on the microarray data. A total of 100 trees were trained; each was trained using a subset of randomly selected instances (genes, 63 %) and variables. Each tree is trained using a subset of the data and is therefore considered a weak classifier. The forest of trees, however, is a strong classifier and is trained using a subset of the data and is therefore considered a weak classifier. The forest of trees, however, is a strong classifier and is trained using a subset of the data and is therefore considered a weak classifier. The forest of trees, however, is a strong classifier and is trained using a subset of the data and is therefore considered a weak classifier. The forest of trees, however, is a strong classifier and is trained using a subset of the data and is therefore considered a weak classifier. The forest of trees, however, is a strong classifier and is trained using a subset of the data and is therefore considered a weak classifier. The forest of trees, however, is a strong classifier and is trained using a subset of the data and is therefore considered a weak classifier. The forest of trees, however, is a strong classifier and is trained using a subset of the data and is therefore considered a weak classifier. The forest of trees, however, is a strong classifier and is trained using a subset of the data and is therefore considered a weak classifier. The forest of trees, however, is a strong classifier and is trained using a subset of the data and is therefore considered a weak classifier. The forest of trees, however, is a strong classifier and is trained using a subset of the data and is therefore considered a weak classifier. The forest of trees, however, is a strong classifier and is trained using a subset of the data and is therefore considered a weak classifier. The forest of trees, however, is a strong classifier and is trained using a subset of the data and is therefore considered a weak classifier. The forest of trees, however, is a strong classifier and is trained using a subset of the data and is therefore considered a weak classifier. The forest of trees, however, is a strong classifier and is trained using a subset of the data and is therefore considered a weak classifier. The forest of trees, however, is a strong classifier and is trained using a subset of the data and is therefore considered a weak classifier.

The RF-generated mean decrease Gini was used as a measure of the importance of a variable (stress condition and time point) for classifying SigB from non-SigB genes. The decrease in the Gini index is determined for each variable at each node. The mean Gini decrease is the sum of all these decreases due to a given variable, normalized by the number of trees in the forest (Breiman, 2001; Klein-Seetharaman et al., 2009; Strobl et al., 2007).

In this analysis, we used two RF models: ‘expression RF’ and ‘kinetic RF’. In the ‘expression RF’ model, gene expression ratios of the wild-type versus the sigB knockout for a given stress and time point were used as variables. The ‘kinetic RF’ model used as variables gene expression ratios of the wild-type strain 168 versus its corresponding 0 min time point for each stress (exponential growth before applying the stress).
Selection of training instances using analysis of variance (ANOVA). In order to derive a stringent training set, the training set was subjected to ANOVA. This additional statistical analysis of the training set was based on the significant differences between the wild-type and the mutant strain in the ‘expression RF’ model, whereas it identifies significant differences between the 0 min time point and the subsequent time points in the ‘kinetic RF model’. GeneSpring GX 7.31 was used to perform one-way ANOVA. This analysis used Welch’s t test, assuming the variances were not equal. The P value cut-off was 0.05 and the Benjamini and Hochberg False Discovery Rate was used for the multiple testing correction.

Promoter prediction. For the identification of promoter sequences upstream of the predicted SigB-dependent genes, a window-based consensus promoter search was conducted using Motif Finder (Decodon). The motif search was performed 300 bp upstream and 100 bp downstream of the CDSs of the B. subtilis genome (Barbe et al., 2009) based on the consensus promoter sequence provided by Petersohn et al. (1999b), with a spacer length of 12–15 bp, allowing two mismatches.

Ranking of genes based on promoter score and expression data. The promoters of the putative SigB-dependent first genes in the operons (127) were aligned, and a consensus sequence for the −10 motif and −35 motif was derived. For each of the motifs, the scores were calculated based on the similarity to the derived consensus, as described by Staden (1984). The overall score of the promoter was the sum of the motif scores of the −10 motif and −35 motif. The resulting scores for all 127 genes were ranked in ascending order.

In addition, the genes were ranked based on expression data. The mean wild-type to mutant ratio of the biological replicates for each condition was calculated, and for each condition, the genes were ranked in ascending order based on the highest expression ratio. Also, for each gene, the mean rank of all these expression ranks for different stresses was calculated.

RESULTS AND DISCUSSION

Our study aimed for a comprehensive description of the structure of the SigB-dependent general stress regulon of B. subtilis. In contrast to previous studies, gene expression was not queried for a single or a few conditions, but most of the conditions described as relevant to SigB activation. These are ethanol, butanol, osmotic and oxidative stress, heat shock, low-temperature growth, and glucose as well as oxygen limitation. To determine the effect of SigB on gene expression, the B. subtilis wild-type strain 168 and its sigB knockout mutant BSM29 were subjected to a comparative analysis under all these conditions in a short time series. The probes in the custom array targeted a total of 350 genes. The selection of genes was based on: (1) 67 genes that were described to be SigB-dependent in all the three previous transcriptional profiling studies (Helmann et al., 2001; Petersohn et al., 2001; Price et al., 2001), (2) 185 genes that were reported to be SigB-dependent in one or two of the three studies, (3) 52 genes reported in other studies to be probably SigB-dependent (Petersohn et al., 1999b, 2001; Price et al., 2001), (4) 10 additional genes of interest based on unpublished data, and (5) 36 genes that were selected as controls as not exhibiting SigB-dependent expression but encoding a variety of other cellular stress-adaptation reactions (for details see Supplementary Table S1).

For all the stress conditions applied, the expression data for both the wild-type and the mutant strains were generated in two independent experiments. For butanol, ethanol, osmotic and oxidative stress, as well as heat shock, a time point immediately before stress exposure and four time points (5, 10, 15 and 20 min) after imposing the stress were analysed. The time points for glucose limitation and oxygen limitation were immediately before and 15, 30, 45, 60 or 90 min after limitation. For the assessment of the gene expression pattern during low-temperature growth at 15 °C, samples were taken during mid-exponential growth at OD540 0.9 and 1.0. In the design of the study we specifically reused stress imposition schemes that had already been reported previously (Antelmann et al., 1996; Helmann et al., 2003; Völker et al., 1995), and decided for environmental stresses to use intermediate stimulus strength that did not completely prevent growth. A graphical representation of the experimental setup with the stress conditions and time points investigated is provided in Supplementary Fig. S1.

To classify ‘SigB’ regulon members and ‘non-SigB’ regulon members we used the RF algorithm, a decision tree-based machine-learning algorithm. Training of the classification model with a training set derived from the previous three studies (Helmann et al., 2001; Petersohn et al., 2001; Price et al., 2001) enabled us to combine the microarray data from our study with prior knowledge. For the group representing true members of the SigB regulon (‘SigB group’), the 67 SigB-dependent genes reported in all three previous genome-wide transcriptome studies (Helmann et al., 2001; Petersohn et al., 2001; Price et al., 2001) were chosen. A set of negative controls (36 genes) derived from genes of different cellular pathways and SigB-independent cellular stress adaptation reactions was selected as the ‘non-SigB’ group. The remaining genes were placed in the test set, and the trained RF model was then used to classify genes in the test set as either ‘SigB’ or ‘non-SigB’ based on the expression profiles.

Here, we report the use of two RF models for the identification of SigB regulon members. The first one was based on the comparison of expression patterns in the wild-type and the sigB mutant, and was thus named ‘expression RF’. However, such a model would likely miss the genes that are subjected to transcriptional control by other regulators in addition to SigB. This additional regulation could trigger a stress-like regulation pattern in the sigB mutant as well, and thereby skew the expression differences between the wild-type and mutant. Thus, the second model, the ‘kinetic RF’ model, was solely based on the expression pattern in the wild-type strain, and was designed to screen for induction by a wide variety of stress stimuli to classify SigB and non-SigB genes.

Defining the SigB regulon using the ‘expression RF’ model

In order to start with a stringent training set and to minimize the number of false positives in the prediction,
the initial training set (67 SigB-dependent genes + 36 SigB-independent controls) was subjected to ANOVA. This analysis did not identify 14 out of the 67 SigB-dependent genes (recognized by all three previous transcriptional profiling studies) as having significantly different expression values between the wild-type and the sigB mutant, and thus these 14 genes (csbC, csbX, dps, mcsA, mcsB, ybyB, ycdF, ycdG, ydbD, ygbX, yjbD, yvqB, yvyD and yxlJ) were removed from the ‘SigB group’ of the training set. Among the 36 control genes, for three genes (spoIIQ, spoIIR and hom) the ANOVA still revealed significantly different expression values between the wild-type and the sigB mutant strain, such that these genes had to be excluded. The remaining 33 control genes were then considered to be the ‘non-SigB’ group of the training set (Fig. 2, Supplementary Table S2). Taking the gene expression ratios between the wild-type and the sigB mutant of the different stress conditions and time points as the 111 variables, the RF classification model was trained based on the training set (53 SigB-dependent genes + 33 SigB-independent control genes) and was subsequently applied to the test set. The out-of-bag (OOB) error of the training was 1.16%, indicating that only a little over 1% of the genes in the training set were incorrectly classified based on

![Diagram](image-url)

**Fig. 2.** Schematic representation of RF analysis. The RF classifier was built based on literature and experimental array data. Both models were trained and used to predict the (non) SigB regulon members. (a) ‘Expression RF’ analysis, where the ratios of gene expression of the wild-type versus the sigB knockout for all stresses were taken as variables. (b) ‘Kinetic RF’ model, where the variables were the ratio of gene expression of the wild-type of each stress to its corresponding 0 min time point. The prediction results (genes classified as SigB or non-SigB) of the two models applied to the test sets are shown. The predicted genes from the two models were compared after including the overlapping genes (67 genes) in each group. The asterisk indicates that the SigB motif was not identified for these seven genes.
the microarray data. Out of the 247 genes of the test set, 99 were predicted to be ‘SigB’ and 148 were classified as ‘non-SigB’ (Fig. 2). Therefore, applying the ‘expression RF’ model resulted in 166 SigB-dependent genes, 67 genes already unequivocally identified by the three transcriptional studies plus 99 genes predicted in this study (Supplementary Table S3).

In addition to the prediction, the RF algorithm provides information about the importance of the variables used (conditions) in the classification model. Based on the expression of the genes forming the training set, the effect on the final classification was assessed for each stress condition. This allowed a ranking of the stress conditions according to their potential to activate SigB (Fig. 3). For the stresses applied, in most cases the first time point after the implementation of stress was the most important for the prediction of SigB dependence. For the environmental stresses, the importance of the further time points decreased with the gradual decline of SigB-dependent expression. Therefore, the previously described transient nature of SigB-dependent genes (Völker et al., 1995) corresponds with the importance of the time points for classifying SigB dependence. Besides the kinetics of SigB induction, a hierarchy of the analysed stress conditions could also be discerned. The importance of butanol stress, osmotic stress and heat shock was higher than that of the other stresses, indicating that in general, SigB-dependent genes responded most strongly during the application of butanol, osmotic and heat stresses (Fig. 3). On the other hand, oxidative stress has only minor importance in the classification of SigB dependence, which corresponds with reports proposing that oxidative stress likely induces only a core set of the SigB regulon (Helmann et al., 2003; Mostertz et al., 2004). The variable importance of the environmental stresses depending on the regulatory phosphatase RsbU was stronger than the variable importance calculated for RsbP-dependent energy stresses (Fig. 3). In addition to the importance of the variables of the training set, we plotted the expression ratios between the wild-type and the sigB mutant of all 152 predicted and confirmed SigB-dependent genes for the time points that displayed maximal induction. The expression-based ranking for ethanol stress was slightly higher when compared with the variable importance-based hierarchy. For all the

![Fig. 3. Variable importance plot resulting from training the ‘expression RF’ model. In this RF model, the variables are the expression ratios of the wild-type to the mutant strain for all the stresses. The bars in the plot represent the mean decrease Gini, an RF-generated variable importance measure. The higher the value of the mean decrease Gini, the more important is the variable for differentiating between ‘SigB’ and ‘non-SigB’ groups. The stress hierarchy in which the order of stress conditions is plotted is based on the maximum variable importance observed. When this importance was related to sensing pathways, the variable importance of RsbU-dependent environmental stresses was higher than that of RsbP-dependent energy stresses. Notes to the figure: 1 butanol stress has not been specifically assigned; we assume that it is quite similar to ethanol stress and assigned it to the RsbU-dependent pathway. 2 Low-temperature growth does not depend on either the RsbU- or the RsbP-sensing pathway (Brigulla et al., 2003). The signalling is also RsbV-independent, but its nature is currently not known, and is therefore not defined (ND). 3 The sensing pathway for oxidative stress is not defined (ND).](http://mic.sgmjournals.org)
other stresses, this expression ratio-based analysis generated a very similar hierarchy of the stress conditions tested (Fig. 4).

**Identification of SigB members with potential secondary regulation using the ‘kinetic RF’ model**

It is known that a proportion of SigB-dependent genes are controlled by secondary regulators as well (Krüger et al., 1996; Petersohn et al., 1999c; Scharf et al., 1998; Varón et al., 1996). These genes are induced under general stress conditions, even in the absence of SigB. For example, clpC has two promoters (SigA and SigB) mapped upstream of the gene. In the wild-type strain the SigB promoter is strongly induced under general stress conditions, whereas the vegetative SigA promoter is weakly induced. In a sigB mutant, the SigA promoter is induced under heat shock and ethanol stress, compensating for the SigB deficiency (Krüger et al., 1996). When only focusing on expression differences between the wild-type and the sigB mutant, the SigB dependency of such genes would likely be missed. Therefore, we performed a kinetic analysis of the wild-type data using the ‘kinetic RF’ model to identify these additional regulon members for which stress-induced expression might also be seen in the sigB mutant strain. Rather than considering induction by individual stresses, we once again exploited the broad activation spectrum of SigB to derive additional SigB regulon members. Here, a stringent training of the model with carefully selected true SigB-dependent genes was of utmost importance. Thus, to ensure the SigB specificity of the candidate genes, a putative SigB promoter, either known from the literature or predicted in our study, was required as an additional criterion to classify genes in the ‘kinetic RF’ model as SigB-dependent.

The training set of the ‘kinetic RF’ model comprised the group of ‘stress induced’ genes and ‘control’ genes. To create a stringent training set, the ‘stress induced’ group, the 67 genes recognized as SigB-dependent by all three previous transcriptional profiling studies, and the ‘control’ group, the 36 selected control genes, were again subjected to an ANOVA based on the wild-type gene expression data. This ANOVA identified 49 genes of the ‘stress induced’ group and 22 genes of the ‘control’ group as displaying significantly different expression after stress imposition compared with exponential growth. For the control group, such expression differences between exponential growth and stress conditions were likely triggered by specific stress regulation, e.g. by osmotic stress for opuAA and opuCA.

![Graph showing expression ratios](image-url)

**Fig. 4.** Box plot of the expression ratios of the wild-type to the mutant of the predicted SigB regulon members (‘SigB’ group from ‘expression RF’ model). For a given stress condition, the time point at which maximum induction was observed was plotted (induction for all the time points for each stress is shown in the box plot in Supplementary Fig. S3). The stress hierarchy observed in this plot with the expression profiles of the predicted SigB-dependent genes (155) was compared with the hierarchy from Fig. 3. The importance of ethanol was slightly higher in the expression box plot than in the variable importance plot; all the other stresses were the same. The maximum induction of genes under ethanol stress was observed at the 10 min time point in the box plot, whereas the importance of the 5 min time point of ethanol stress is still higher than that of the 10 min time point in the variable importance plot (Fig. 3).
After removal of 18 genes from the ‘stress induced’ group and 22 genes from the ‘control’ group, the ‘kinetic RF’ model was trained with 49 ‘stress induced’ and 14 ‘control’ genes. This trained model was then used to classify the test set of 247 genes based on the expression pattern in the wild-type using 92 variables (the above-mentioned ratios to time point 0 of the wild-type strain). The OOB error of training the ‘kinetic RF’ was 12.7%. This OOB error is higher than that of the ‘expression RF’ (1.16%). However, an OOB error of 12.7% is still acceptable, indicating that the trained RF model can still classify the majority of SigB members as distinct from the non-SigB members.

A total of 121 genes were predicted to be ‘stress induced’ and 126 genes were predicted to be part of the ‘control’ group (Fig. 2). Thus, 188 genes (67 genes already predicted by the three transcriptional profiling studies and 121 genes predicted from the test set) were classified as induced by diverse stresses in the wild-type strain.

**Integrating the two RF-based predictions**

The comparison of the results of the ‘expression RF’ and ‘kinetic RF’ models revealed a large overlap of 151 genes (Fig. 2). Besides these 151 genes, 37 additional genes were shown to display strong induction by multiple stresses in the wild-type strain by the ‘kinetic RF model’ (Fig. 2). These genes still showed considerable expression in the sigB mutant strain and were therefore not detected as SigB-dependent in the ‘expression RF’ model, indicating control by additional regulators. For 20 of the 37 genes, additional regulation by secondary regulators has already been proven by previous studies. These 20 genes are: bioB, bmrR, purK, radA, yacK, yacM, ydaF, ykrT, yocL, ypuD, yraA, ysnF, ywH, yxG and the operon yceC-ycE-ycF-ycG-ycH (Au et al., 2005; Banse et al., 2008; Baranova et al., 1999; Bower et al., 1996; Cao et al., 2002; Chu et al., 2006; Chumsakul et al., 2011; Comella & Grossman, 2005; Dreweicki et al., 1998; Ebbole & Zalkin, 1987; Eiamphungporn & Helmann, 2008; Erwin et al., 2005; Grundy & Henkin, 1998; Höper et al., 2005; Jervis et al., 2007; Kearns et al., 2005; Krüger et al., 1996; Kumaraswami et al., 2010; Leelakriangsak et al., 2007; Lei et al., 2009; Marvasi et al., 2003; Nakano et al., 2003; Nguyen et al., 2009; Perkins et al., 1996; Petersohn et al., 2001; Sekowska & Danchin, 2002; Wang et al., 2006; Weng et al., 1995; You et al., 2008). Details about their secondary regulators are included in Supplementary Table S4. The remaining 17 genes, nap, opuBA, opuBB, ydB, ydJL, yfMG, yhaS, yhaT, yhaU, yisP, yggE, yqY, yqXH, ytzE, yvaC, ywAF and ywM, identified in our study are also potentially controlled by a secondary regulator.

As these 37 candidate genes were predicted solely because of their stress induction in the wild-type, further evidence

---

**Fig. 5.** Hierarchical clustering of genes based on the expression ratio ranks of the individual stress conditions after Z-scale transformation (a). Hierarchical cluster with indicated cluster C1–C4 of the mean of all expression ratio ranks and the corresponding promoter scores (b). For both analyses, of the 179 genes having a predicted SigB-dependent promoter, only the first genes in possible operons were considered. The expression ratios and promoter scores for the resulting 127 first genes were ranked in ascending order as described in Methods. Higher expression ratios, as well as the higher promoter scores, got the lowest ranks and thereby obtained the lowest Z-scores (coloured in red). Lower expression ratios and low promoter scores had higher ranks and received higher Z-scores (coloured in green).
of their SigB dependence was required. We required the presence of a SigB motif either identified in a window-based DNA motif search or reported previously (Helmann et al., 2001; Petersohn et al., 2001; Sierra et al., 2008). Since seven (nap, ykrT, yocL, yagE, yteE, ywaf and ywhH) out of the 37 genes failed to display such a SigB motif, only the remaining 30 genes were considered to indeed possess SigB-dependent regulation (Supplementary Table S4). Overall, including the remaining 30 genes, 196 SigB-dependent genes were identified in our study (Fig. 2). Of these 196 SigB regulon members, 179 genes have been assigned to a putative SigB-dependent promoter either in the above-mentioned previous studies or in the current study (Supplementary Table S5).

In order to compare the promoter strength with the expression ratios of the different stresses the expression ratios were ranked for each stress condition (for details see Methods) and hierarchically clustered after Z-score transformation (Fig. 5a). As the ranks for each gene behaved quite similarly for all stresses, the mean of all expression ranks for each gene was calculated. The predicted SigB-dependent promoters were scored based on their similarity to the derived consensus sequence and ranked accordingly. These promoter ranks and the corresponding mean expression ranks were clustered and represented in a heat map (Fig. 5b). The clustering resulted in four clusters (C1–C4). The two major clusters, C1 and C4, harbour genes having good promoter scores (promoter similar to the consensus) and high expression ratios, and genes having low promoter scores and low expression ratios, respectively. As expected, the majority of the experimentally verified promoters can be found in cluster C1. Cluster C2 contains genes having good promoter scores and lower expression ratios. Six of the 20 genes within cluster C2 are subject to additional regulation according to SubtiWiki (Flórez et al., 2009). Cluster C3 consists of genes having low promoter scores and high expression ratios. Cluster C3 together with C4 contains fewer genes that have an experimentally verified promoter compared with the remaining clusters. Based on this analysis we conclude that there is a correlation between expression ranks and promoter scores for the genes in clusters C1 and C4. A detailed list of the clusters is provided in Supplementary Table S6.

The list of 196 genes predicted in our study is also supported by the results of an extensive phenotypic screening performed by Höper et al. (2005). That study identified the stress sensitivity of 94 mutants of candidate SigB regulon members using phenotypic screening of the corresponding mutants under severe stress conditions such as ethanol stress, heat shock, salt stress and low-temperature growth. Of the 94 mutants analysed, 80 were sensitive to at least one of the four stress conditions tested. In the current study, 74 of the 80 genes encoding these functionally important proteins were studied and 92% of them were verified as SigB-dependent. According to the data of Höper et al. (2005), sodA, yceC and yqjL mutants showed the strongest differences from the wild-type during low-temperature growth. In our analysis, these genes were strongly induced not only under low-temperature growth but also under all the conditions tested. This example also highlights the value of our study, because these genes were previously identified only by Petersohn et al. (2001), and would have been missed if the focus had been on the 67 overlapping genes.

Around 95% of genes which were described as being SigB-dependent in at least two of the three previous transcriptional profiling studies (Helmann et al., 2001; Petersohn et al., 2001; Price et al., 2001) were validated in the present study (Fig. 6). However, for genes reported to be SigB-dependent in only one of the three studies the frequency of validation in our study was clearly lower. Of the 96
candidate genes identified only by Price et al. (2001), 62 were not identified as SigB-dependent in our study (Fig. 6). Surprisingly 60% of these genes have a predicted SigB promoter, despite the fact that they did not show induction under most of the tested stresses. Thus, the specific induction of most of these genes (90%) only after ethanol stress likely reflects a thus-far-unidentified ethanol-specific regulation deviating from the typical SigB-dependent induction by other stimuli.

Concluding remarks

We present a systematic identification of SigB regulon members using two RF models, ‘expression RF’ and ‘kinetic RF’, based on a comprehensive microarray study in which most of the SigB-inducing stress conditions were queried. The targeted analysis used in this study covered all potential SigB-dependent candidate genes based on previous knowledge. Inclusion of environmental and energy stresses as well as low-temperature growth also ensured that prospective SigB targets displayed a broad induction pattern, which should be an important determinant in assigning a gene to the regulon. Integrating the observations of previous studies with those of the current study we propose 196 SigB regulon members. This knowledge can serve in further studies to reveal the specific functions of these genes in B. subtilis stress management.

ACKNOWLEDGEMENTS

We thank Jörg Bernhardt for providing the Motif Finder software. This work was supported by the transnational BaCell-SysMO project.

REFERENCES


Edited by: J. M. van Dijl