New insights into the regulatory networks of *Bacillus subtilis* controlling general stress response and osmo adaptation

I n a u g u r a l d i s s e r t a t i o n

Zur

Erlangung des akademischen Grades

Doktor rerum naturalium (Dr. rer. nat.) an der Mathematisch-

Naturwissenschaftliche Fakultät

der

Ernst-Moritz-Arndt-Universität Greifswald

Vorgelegt von

Priyanka Nannapaneni

geboren am 15.08.1985

in Kodali – India

Greifswald, August 2012

Dekan: Prof. Dr. rer. nat. Klaus Fesser

1. Gutachter: Prof. Dr. rer. nat. Uwe Völker

2. Gutachter: Prof. Dr. rer. nat. Jörg Stülke

Tag der Promotion: 07. 01. 2013

TABLE OF CONTENTS

SUMMARY

The soil living, Gram-positive bacterium *Bacillus subtilis* is frequently exposed to a wide variety of stress and starvation conditions in its natural environment. In order to survive under these environmental and energy stresses, the bacterium acquired a general stress response mechanism mediated by the alternative sigma factor, SigB. A wide-variety of stress conditions such as environmental stress conditions like ethanol stress, heat stress, oxidative stress, osmotic stress or limitation of glucose, oxygen, phosphate etc.; and low temperature growth induce this SigB-dependent general stress response. Though much is known about the mechanisms of activation of this general stress response, the conditions that induce the SigB regulon and its general functions, the definition of the structure of the SigB regulon is not completely clear. The SigB-dependent general stress regulon has previously been characterized by proteomic approaches as well as DNA-array based expression studies. Genome-wide expression studies performed by Price, Petersohn and Helmann defined the SigB regulon containing well above 100 target genes, however the overlapping list of target genes contains only 67 members. The differences between these studies probably result from the different strains, growth conditions, array platforms and experimental setups used in these studies.

The first part of this work presents a targeted microarray analysis, which was performed to gain a better understanding of the structure of the general stress regulon. This is the first study analyzing the gene expression of a wild type strain and its isogenic *sigB* mutant strain for almost all known SigB inducing conditions, using the same array platform. Furthermore, the kinetics of the gene expression of 252 putative SigB-dependent genes and 36 appropriate control genes were recorded. The data were analyzed using Random Forest, a machine-learning algorithm, by incorporating the knowledge of previous studies. Two Random Forest models were designed in this study. The "expression RF" model was designed to identify genes showing expression differences between wild type and *sigB* mutant and the "kinetic RF" model to identify genes having a SigB-dependent expression kinetic, but is subject to secondary regulators next to SigB influencing their expression in the *sigB* mutant.

The random forest classification using the "expression RF" model identified 166 genes as SigB regulon members based on the expression differences between the wild type and the *sigB* mutant strain. A variable importance plot showing the impact during the classification process within the "expression RF" could assign a hierarchy to the stress conditions investigated in this study. This hierarchy suggested all the RsbU-dependent environmental stresses to have higher impact on SigB-dependent gene expression compared to the RsbP-dependent energy stresses. The "kinetic RF" model identified 30 additional genes, having additional regulators next to SigB. The SigB dependency of the 30 genes identified by the "kinetic RF" model was validated by screening for SigB promoter motifs within the upstream region of these genes. The hierarchical clustering of the obtained motifs scores with the expression ratios of the SigB regulon members predicted in the current work revealed that only a subset of genes displayed correlation of gene expression values and sequence motifs. As this observation is not true for all sets of genes, it cannot be generalized that gene expression is only correlated with the corresponding motif scores.

In total 196 SigB regulon members could be classified by this targeted oligo nucleotide microarray study. The majority of these regulon members were preceded by a putative SigB promoter motif either identified previously or predicted in the current work. The inclusion of the broad range of stress conditions, from environmental stresses to energy limiting conditions enabled a more detailed characterization of the structure of the general stress regulon of *B. subtilis.* The implementation of machine learning algorithms allowed the prediction with a minimum number of false-positives.

In the second part of this work a high resolution tiling array analysis for the majority of growth conditions, stresses and changes in carbon sources supply was exploited for the screening for new SigB targets within already annotated or newly annotated RNA features. Thereby 133 previously un-annotated RNA features, which were completely new, were assigned to the SigB regulon. 50 of these 133 new features encode antisense RNAs which can have potential influence on the transcription / translation of their sense RNAs targets. A set of 282 annotated genes were indentified to be SigB regulon members, comparison with the targeted oligo nucleotide study, 90 genes were newly identified and not known to be SigB-dependent before. The analysis of the expression levels of these genes by k-means clustering revealed a cluster of 32 genes having low

induction levels in all SigB-inducing conditions, although the majority of these genes possess a well-conserved SigB promoter motif. However, all these genes are probably subject to the control of regulators other than SigB, which might mask the typical strong SigB-dependent induction in the analyzed stress conditions.

The analysis of the expression levels of the SigB regulon under a variety of conditions, revealed the SigB-dependent expression in conditions such as growth on plates, in swarming cells, biofilm formation and growth on glycerol as a carbon source. The possible reason for the induction of the SigB regulon during growth on plates and in swarming cells was supposed to be due to scarcity of the nutrients on plates, e.g. glucose limitation. SigB-dependent genes were likely induced during growth on glycerol due to the oxygen limitation that arose during the growth. However, induction of the SigB regulon during biofilm formation is assumed to be due to the phosphate limitation. The description of these new SigB activating stimuli gains support from the fact that the majority of the SigB-dependent genes were induced under these growth conditions.

In addition to the general stress response, *B. subtilis* cells have stress specific adaptive mechanisms such as osmotic response, which was addressed in the third part of this dissertation. The frequent flooding and drying of the soil triggered osmotic stress, one of the most common stress conditions encountered by soil bacteria. Bacterial cells are equipped with osmo-specific adaptation responses in which specific regulation of a set of genes is used to maintain proper cellular function. It was known from previous studies that a large set of genes were influenced in expression by salt shock as well as growth at high osmolarity. Detailed analysis of the tiling array data revealed 467 differentially regulated newly annotated features during salt shock and 251 newly annotated features that were expressed at a different level during continuous growth at high versus low osmolarity.

A comparison of the studies that used the *sigB* knockout mutant with the tiling array study also provided support for the sigma factor competition in control of the expression of osmo-adaptive genes. The level of induction of specific osmo-adaptive genes was much higher in the *sigB* mutant strain compared to the wild type strain. Furthermore, the tiling array data revealed a SigB-dependent antisense RNA *S1290* upstream of the *opuB* operon that transports choline to the cell. The presence of this

antisense RNA had a potential impact on the transcription of the *opuB* operon, during salt shock.

In agreement with the previous studies, the tiling array data assigned the osmotically regulated *proHJ* operon to the SigE regulon, with a SigE promoter upstream. In addition, the significantly higher percentage of proline among spore coat proteins also supports the assumption that osmotic synthesis of proline might play a role during the generation of spores.

In conclusion, the tiling array data revealed newly annotated RNA features that are regulated during the general stress response as well as the osmotic response of the cell. The current work identifies new conditions that induce the majority of SigB-dependent genes as well as the new features that regulate the osmotically induced genes.

1. INTRODUCTION

1.1 STRESS RESPONSES OF *BACILLUS SUBTILIS*

The Gram-positive *Bacillus subtilis*, originally named as *Vibrio subtilis* (Ehrenberg, 1835), is one of the most commonly found soil living bacteria. The extensive studies of sporulation, competence (the uptake of exogenous DNA), motility, the expression of degradative enzymes, and antibiotic production in *B. subtilis* made the organism a laboratory model organism for its class, the Gram-positive bacteria. Moreover, it is one of the most advanced microorganisms for systems biology approaches (Mäder *et al.*, 2012). As an organism living in the upper layer of soil, *B. subtilis* is often exposed to different growth-limiting environmental conditions, such as changes in temperature and osmolarity as well as nutrient limitation. For the survival and proliferation under this wide range of stress conditions, the general stress responses of *B. subtilis* mediated by the alternative sigma factor SigB plays a critical role. Besides this general stress response, each stress triggers the activation of stress specific proteins. For example, during osmotic stress the cells accumulate compatible solutes like proline, glycine betaine and choline through osmotically regulated transporters and synthesis. Therefore, proteins like ProHJ, OpuAA-AB-AC, OpuBA-BB-BC-BD, OpuCA-CB-CC-CD, OpuD, and OpuE were produced at higher levels (Boch *et al.*, 1994; Bremer, 2002). After heat stress, the expression of the chaperone operons like *dnaK* and *groESL,* which are classified as class I heat shock genes, is driven by the vegetative promoter SigA (Yuan & Wong, 1995). During oxidative stress, the peroxide (PerR) regulon is induced encoding proteins like PerR, AhpC, AhpF, KatA, MrgA, Spx, HemA-X etc (Bsat *et al.*, 1998; Chen *et al.*, 1995). Some of these proteins like AhpC were also slightly induced after heat shock and glucose limitation. Thus, the cells have their specific set of proteins activated during particular stresses. In case of nutrient starvation, to mediate the adaptation of cellular functions, global regulatory systems governed by RelA, CodY, σB, σH were activated (Mäder *et al.*, 2007). The stress specific responses can be pathway specific like activation of CcpA and TrnA, the regulators of central carbon and nitrogen metabolism, respectively. Additionally, the cells also possess some long-term adaption responses like competence, chemotaxis, stringent response, secretion of exoenzymes to utilize alternative nutrients and endospore formation. The formation of highly resistant

endospore is a time and energy consuming process. It is induced by nutrient limitation and initiated via the phosphorylation of the Spo0A response regulator (Grossman, 1991; Losick *et al.*, 1986). The phosphorylation of Spo0A is concerted by a multicomponent phosphorelay wherein the sequential transfer of phosphate between four proteins, kinases (major kinase: KinA), Spo0F, Spo0B and Spo0A (Burbulys *et al.*, 1991). The initiation of sporulation depends on the amount of phosporylated Spo0A where in lower levels of phosphorylated Spo0A results in the development of biofilms and cannibalism (Higgins & Dworkin, 2012).

Phosphorylated Spo0A controls expression of genes required for the asymmetric division of the cell, as well as genes encoding SigF and SigE regulators (Errington, 2003; Hilbert & Piggot, 2004; Steil *et al.*, 2005; Stragier & Losick, 1996). The rod-shaped cell is partitioned into a small forespore and a large mother cell. The mother cell engulfs the forespore to initiate spore metabolic dormancy and the synthesis of spore DNA. The forespore is converted to a spore through the production of the spore cortex, inner and outer coat. The mother cell then lyses and releases the mature spore. The entire process of sporulation involves five sporulation specific sigma factors and at least five other transcription factors. The sigma factor SigF controls the gene expression during the early stages of fore spore development wherein during the later stages of post-engulfment SigG replaces it (Errington, 2003; Hilbert & Piggot, 2004; Stragier & Losick, 1996). The sequential activation of the two mother cell specific sigma factors, SigE and SigK, mediates the early and late stages of gene expression within the mother cell. These two sigma factors initiate the expression of genes encoding spore coat proteins, which are additionally regulated by the transcription factors SpoIIID and GerE, respectively (Losick & Pero, 1981; Piggot & Losick, 2002). After the initiation of septation, the process is irreversible and the cells cannot stop sporulating. Thus, many times this is the last choice, when the conditions become completely unfavorable.

The current study deals with the non-specific general stress response via the alternative sigma factor SigB and the specific osmotic response, and these are described in more detail in the following parts of the thesis.

1.2 SIGB REGULON OF *BACILLUS SUBTILIS*

The alternative sigma factor SigB was the first sigma factor identified in *B. subtilis* by Haldenwang and Losick (Haldenwang & Losick, 1980). The activation of SigB controls gene expression during stress conditions in exponentially growing as well as in starving non-growing cells. Increased expression of the SigB regulon provides *B. subtilis* cells with a non-specific, multiple stress resistance against the wide range of stresses (Gaidenko & Price, 1998; Höper *et al.*, 2005; Völker *et al.*, 1999). This SigB-dependent gene expression also contributes to growth and survival at low temperatures and during extended stationary phases (Brigulla *et al.*, 2003; Mendez *et al.*, 2004).

The collection of SigB-activating stress conditions includes sudden environmental stresses like heat-, acid-, ethanol-, butanol and osmotic stress, Mn^{2+}, nitric oxide (NO), sodium nitroprusside (SNP), red and blue light irradiation; energy stresses including entry into stationary phase, 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 extremely low temperatures (Avila-Pérez *et al.*, 2006; Avila-Pérez *et al.*, 2010; Brigulla *et al.*, 2003; Hecker *et al.*, 2007; Hulett, 2002; Price, 2000; Price, 2011; Völker *et al.*, 1995). These inducing conditions can be broadly classified into three classes: environmental, energy stresses and low temperature growth. Separate signal transduction pathways signal these three classes.
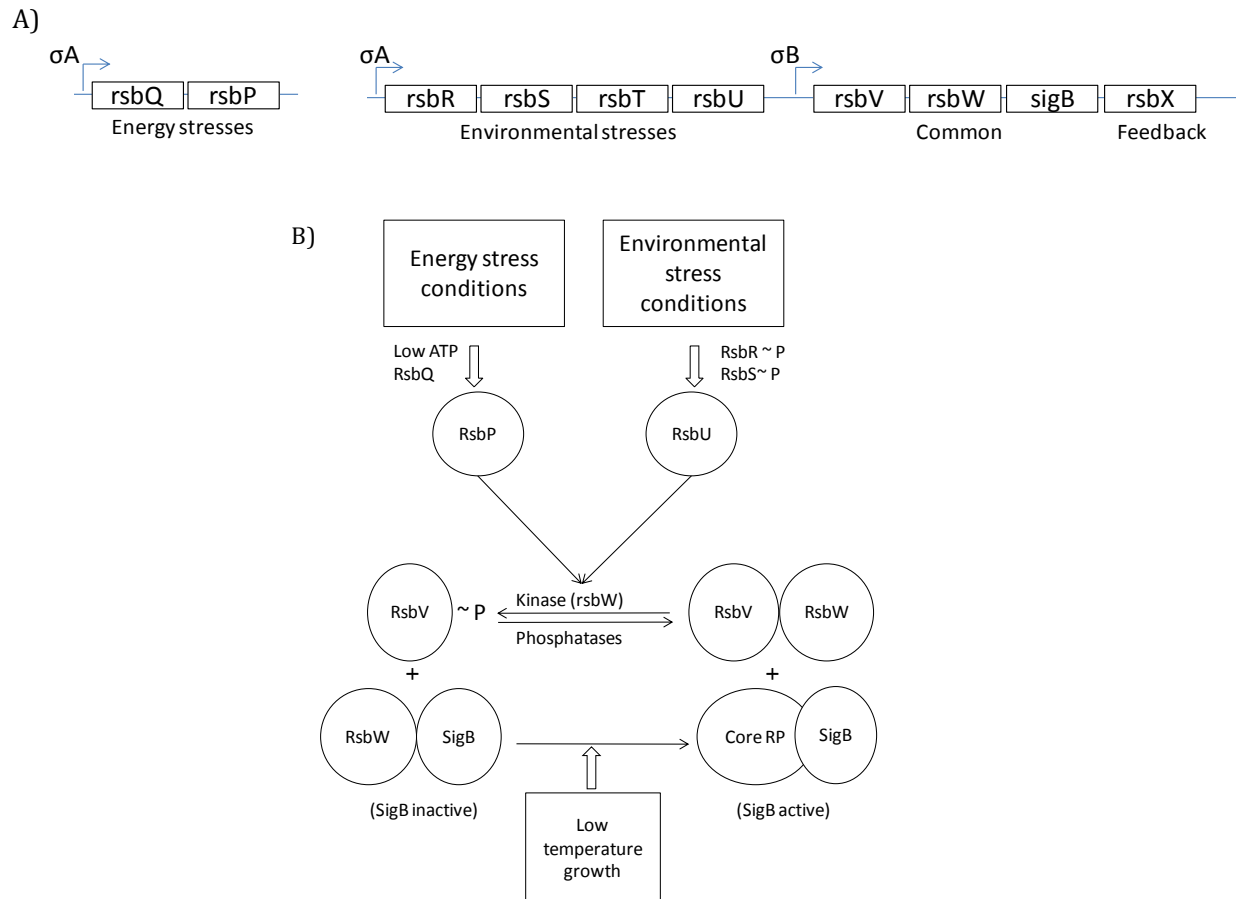
Fig. 1: The regulatory network of SigB in *B. subtilis.* (A) Graphical representation of the *sigB* operon and the operons coding for the regulatory proteins of SigB. (B) The schematic representation of the signal transduction pathways that control the activity of SigB. The entire environmental stress conditions signal via RsbU and all the energy stresses signal via RsbP. Both signaling pathways lead to dephosphorylation of RsbV, thus binding to RsbW. This leaves SigB free to bind to RNA polymerase and initiate SigB-dependent transcription. For low temperature growth the signaling is still unknown but it could be shown that it is RsbV independent (Brigulla *et al.*, 2003). (Figure adapted from Hecker *et al.* 2007, Price *et al.* 2011)

This whole signaling network functions by a partner-switching mechanism, which includes the target protein σ^B, the anti-sigma factor RsbW and its antagonist protein RsbV in which key protein-protein interactions are controlled by serine phosphorylation and dephosphorylation. The eight-gene *sigB* operon consists of *sigB* and seven *rsb* genes (shown in Fig. 1A). The entire operon is under the control of a SigA promoter whereas an internal SigB promoter is responsible for the autocatalytic induction of the four downstream genes of the operon. Under normal conditions, the phosphorylated anti-anti sigma factor RsbV is unable to bind the anti sigma factor RsbW. The RsbW protein is then free to bind SigB and form a stable complex thereby prohibiting the binding of SigB to RNA polymerase (Alper *et al.*, 1996; Benson &

Haldenwang, 1993; Dufour & Haldenwang, 1994). The activation of SigB depends on the release of RsbW by dephosphorylating RsbV. Thus, RsbV and RsbW are the positive and negative regulators of SigB, respectively, and are often termed as the primary regulators of σ^B activity (Hecker & Völker, 2001; Hecker *et al.*, 2007; Price, 2000).

The environmental stress conditions trigger the signaling via the PP2C-type phosphatase RsbU involving at least 10 proteins. This signaling leads to the dephosphorylation of RsbV and subsequent activation of SigB (Fig. 1B). The activation of SigB is not prolonged and the transient induction of the SigB activity during environmental stresses is probably because of the negative feedback loop of RsbX (Igoshin *et al.*, 2007), which leads to the phosphorylation of RsbS and RsbR (Akbar *et al.*, 1997; Kang *et al.*, 1996; Völker *et al.*, 1995; Yang *et al.*, 1996). The second class of stress conditions, energy stresses, is also signaled by a PP2C-type phosphatase, RsbP. During energy limitation, RsbP is activated and dephosphorylates RsbV. This leads to the activation of SigB (Vijay *et al.*, 2000; Völker *et al.*, 1995). Despite these two signaling pathways, there is a third signaling pathway during the growth at low temperatures. The activation under this stress condition is independent of RsbV and the mechanism of activation is not well characterized (Brigulla *et al.*, 2003).

The *rsbU* and *rsbP* single mutant strains abolish the SigB activity under their respective class of activating stress conditions, whereas no effect is observed on the other stress class. The *rsbW* mutants have a severe growth defect, arising from the unrestricted high expression of SigB. The *sigB* mutant is viable and does not exhibit any growth defect under standard growth conditions. The s*igB* mutant cells displayed an at least 50 to 100-fold decrease in the survival to various stress conditions (Völker *et al.*, 1999). Furthermore, when pre-adapted with mild stress, show an impaired survival after the imposition of harsh stresses compared to the wild type strain (Völker *et al.*, 1999).

The general stress response of *B. subtilis* has been extensively studied since the 1990's. Several groups identified regulon members by various approaches including transposon mutagenesis (Boylan *et al.*, 1991; Boylan *et al.*, 1992; Boylan *et al.*, 1993), gel-based proteome analysis or a 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 by three genome-wide transcriptional profiling studies (Helmann *et al.*, 2001; Petersohn *et al.*, 2001; Price *et al.*, 2001).

The three major DNA array studies identified SigB regulon members by testing the 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 and a *sigB* mutant strain cultivated in complex LB medium and postulated 193 genes to be SigB-dependent. Helmann *et al.* (2001) have conducted experiments exposing a wild type strain to heat shock (shift from 37°C to 48°C) in minimal medium, using DNA microarray glass slides. This study identified 123 genes as members of the SigB regulon. Petersohn *et al.* (2001) have comparatively investigated a wild type strain and its isogenic *sigB* mutant after ethanol stress (4% v/v), heat shock (shift from 37°C to 48°C) and osmotic stress (4% w/v NaCl to 10% w/v NaCl). They assayed the transcriptional patterns using DNA macroarrays containing PCR products of the whole genome and 124 genes were claimed to be SigB-dependent.

All these three studies (Helmann *et al.*, 2001; Petersohn *et al.*, 2001; Price *et al.*, 2001) postulated rather different sets of SigB regulon members. The overlap among the identifications can be seen in Fig. 2. When comparing these different sets of SigB regulon members, it became obvious that the overlap was rather small. Out of the 261 genes postulated in these studies, only 67 genes were described to be SigB-dependent in all three studies. The potential reasons for the variety between the identified sets 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.

Price *et al.* (2001)   Petersohn *et al.* (2001)
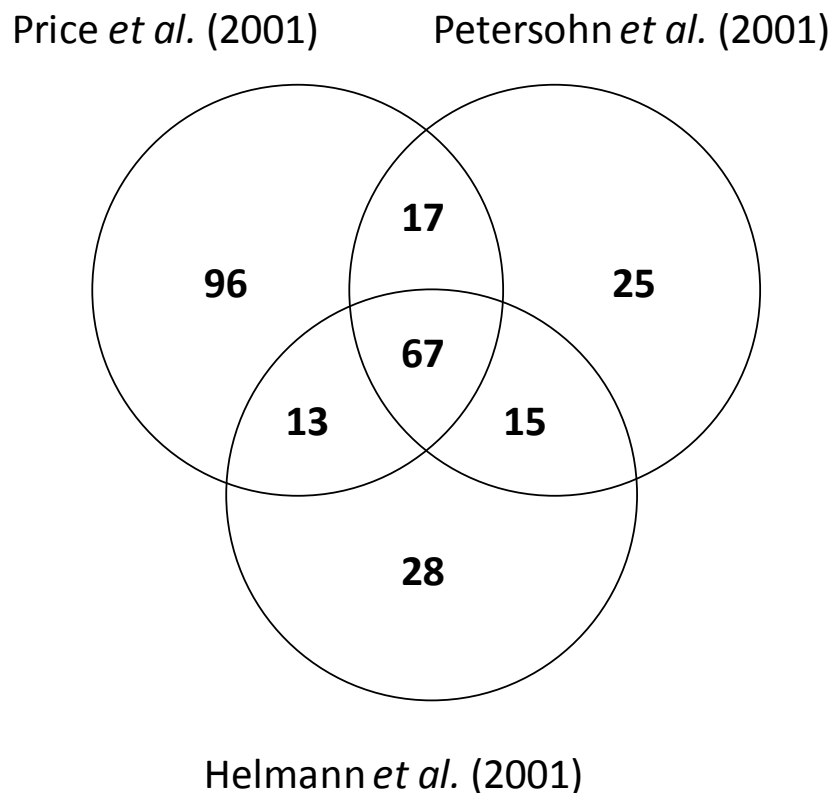
Helmann *et al.* (2001)

Fig. 2 Venn Diagram comparing the gene lists of SigB-dependent genes identified in the previous transcriptome studies (Helmann *et al.*, 2001; Petersohn *et al.*, 2001; Price *et al.*, 2001) Overlapping areas denote genes recognized by two or more studies, and those in the outer areas represent genes identified in individual studies.

Among the identified SigB genes by any of these studies, only a small proportion of around 30 genes code for the proteins with known function (Hecker *et al.*, 2007; Hulett, 2002). Some examples include the proteins involved in oxidative stress resistance (KatE, KatX, Dps, OhrB, TrxA), cell wall biosynthesis (GtaB), osmotic stress resistance (OpuE), heat stress resistance (ClpC, ClpP), antibiotic resistance (BmrU, BmrR) etc. 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. This is evident from a study of Höper *et al.* (2005), where the authors conducted 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. These single mutants were individually subjected to ethanol stress (10% v/v), osmotic stress (10% w/v), heat stress (54°C) and low temperature growth (survival at 4°C and growth at 12°C) (Höper *et al.*, 2005). This analysis revealed that 85% of the

mutants carrying knockouts in single SigB-dependent genes displayed increased stress sensitivity to at least one of the stress conditions tested. Detailed proteome analysis along with BLAST homology searches for one of the candidates, YerD, gave reasons to assume a significant role in synthesis of glutamate as a precursor for proline during osmotic stress response. However, a more comprehensive screening of the function of SigB-dependent genes requires precise knowledge of the structure of the SigB regulon.

The structure of the SigB regulon can thus be elucidated from the analysis of a comprehensive targeted comparative microarray study of the expression changes observed in the *B. subtilis* wild type strain and its isogenic *sigB* mutant using all the possible SigB inducing stress conditions. However, a part of SigB-dependent genes was reported to be controlled by secondary regulators as well (Krüger *et al.*, 1996; Petersohn *et al.*, 1999a; Scharf *et al.*, 1998; Varon *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 the SigB deficiency. Thus, while elucidating the structure of SigB regulon, these additional regulators should be taken into account.

1.3 OSMOTIC RESPONSE OF *BACILLUS SUBTILIS*

As an organism residing in the upper layer of soil, along with other different kinds of stresses, *B. subtilis* is often exposed to osmotic stress. This is due to the frequent drying and flooding of the soil. Thus, due to the sudden change in the concentration of salt around the cell, the cell needs to adopt a specific mechanism to cope up the stress. The initial stress response to high osmolarity is the uptake of potassium (K^+). This uptake is mediated by turgor-responsive transport systems (KtrAB and KtrAC) and is essential to counteract the immediate outflow of water from the cell after the shock and the subsequent resumption of growth (Whatmore *et al.*, 1990; Whatmore & Reed, 1990) (shown in Fig. 3).

The uptake of K$^+$ is followed by the accumulation of compatible solutes in the cell. The compatible solutes comprises amino acids and its derivatives, sugars and polyols, quaternary amines and their sulphonium analogues, sulphate esters, N-acetylated diamino acids, small peptides etc (Csonka, 1989). They act like chemical chaperons, stabilizing the enzymes and cell components against the denaturing effects of high ionic strength (Potts, 1994). Thus, these compatible solutes, which enhance the growth of the cell under high osmolar concentrations protecting the cell, can act as osmoprotectants.

B. subtilis cells have five osmotically regulated transport systems for the uptake of osmoprotectants from the environment. Among these, OpuA (OpuAA, OpuAB, OpuAC) and OpuC (OpuCA, OpuCB, OpuCC, OpuCD) are multi-component ABC-type transporters. In terms of substrate specificity, the OpuC system has a broader substrate spectrum and the wide range of osmoprotectants transported by this system is shown in Fig. 3. The multi-component ABC-type transporter OpuB (OpuBA, OpuBB, OpuBC, OpuBD) is a highly specific transporter that transports choline into the cell, which is subsequently converted into glycine betaine in the cell. OpuD and OpuE are single component transporters and are highly substrate specific. The earlier transports only glycine betaine into the cell and the later is a high affinity proline transporter. *B. subtilis* cells transport proline from the plant root exudates and decaying plant material in the natural habitat and from the culture medium *in vitro*. Both transport systems are known to be induced by SigB-dependent general stress response, induced upon several stress and starvation conditions.
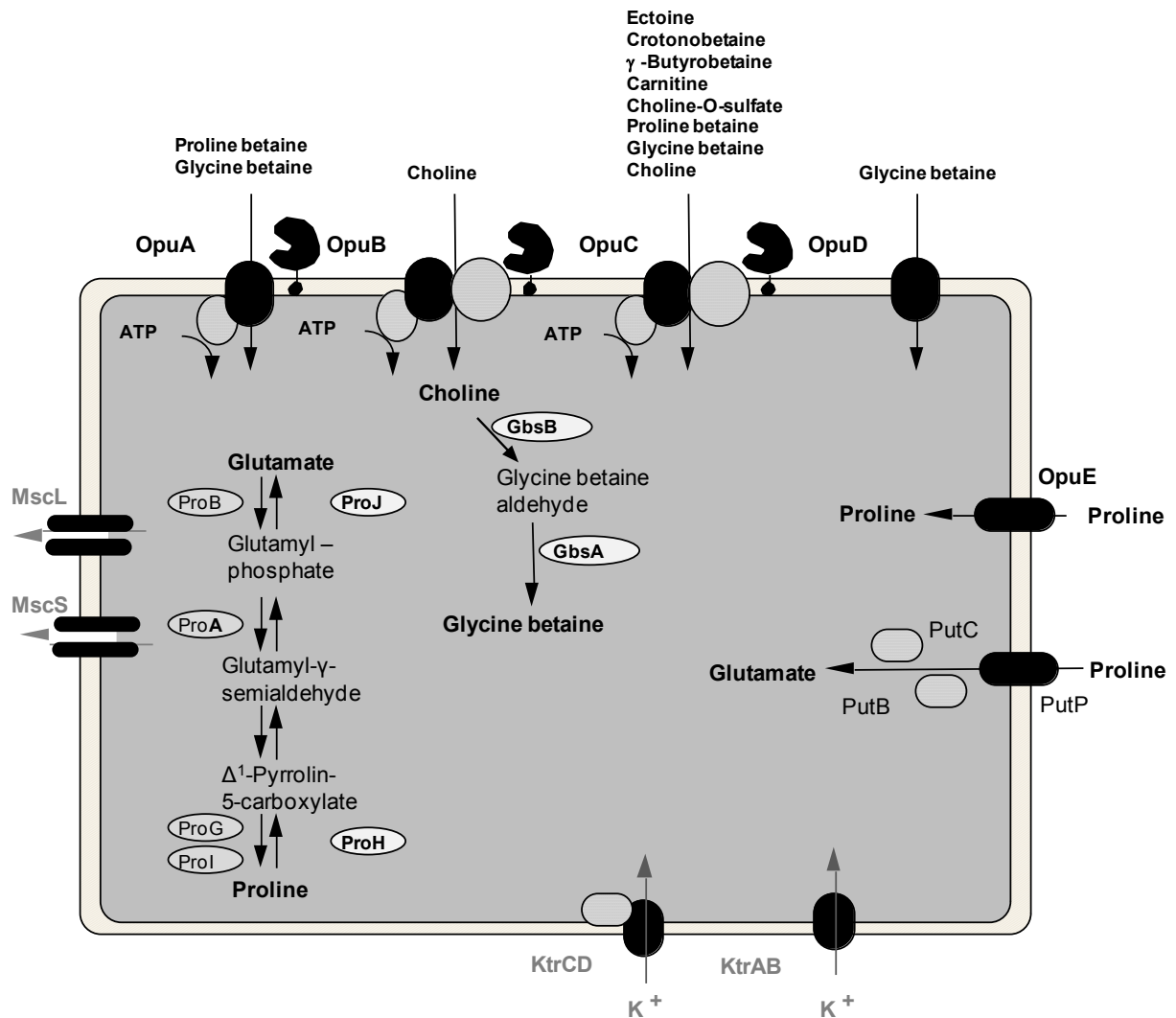
Fig. 3 Osmotic stress response in *Bacillus subtilis.* The accumulation of osmoprotectants like proline, glycine betaine, choline etc. is realized either by transport via the Opu transporters or, in case of glycine betaine and proline, synthesis within the cell (Bremer, 2002; Steil, 2004).

In addition to the uptake of osmoprotectants from the environment, proline and glycine betaine are synthesized in the cell. During normal growth conditions, anabolic synthesis of proline is controlled through a feedback inhibition of the ProB-encoded γ-glutamyl kinase. When the cells are exposed to high osmolar conditions, the cells needs to synthesize very large quantities of proline using a distinct set of proteins for the biosynthesis, mediated by ProH (a pyrroline-5-carboxylate reductase) and ProJ (a glutamate 5-kinase). This osmotically controlled synthesis of proline is critical for the cellular defense against high osmolarity. Both these osmo-adaptive and anabolic proline biosynthetic routes are interlined via the γ-glutamyl phosphate reductase (ProA). Glycine betaine is mainly synthesized by the cell from its precursor choline during the

deficit of water in the surroundings (Boch *et al.*, 1994). The soluble NAD dependent dehydrogenases GbsB and GbsA are used in the two step conversion of choline to glycine betaine with an intermediate product glycine betaine aldehyde. The transcription of the operon *gbsA-gbsB* is enhanced under the presence of choline (Boch *et al.*, 1994). These genes (*opuAA-AB-AC, opuBA-BB-BC-BD, opuCA-CB-CC-CD, opuD, opuE, proHJ, gbsAB)* are the osmo-adaptive genes where the osmotic response in the cell triggers the induction/repression of a specific set of genes.

During low osmotic conditions, these osmoprotectants are a threat to the integrity of cell. Mechanosensitive channels (MscL, MscS) located on the cell membrane are responsible for the rapid release of compatible solutes and ions.

Upon shock, along with the osmo specific responses in the cell, the SigB-dependent general stress response is activated. The ECF (Extra Cytoplasmic Function) sigma factors SigW and SigM are also subsequently induced under high osmolar conditions. Steil *et al.* (2003) investigated the genome-wide transcriptional profiling of *B. subtilis* in a *sigB* mutant, under the conditions of high osmolar growth and osmotic shock. This study revealed 126 genes differentially regulated under salt shock, 75 genes induced and 51 genes repressed whereas the induction of 123 genes and repression of 101 genes during continuous growth at high osmolarity. Recent studies of Hahne *et al.* (2010) reveal the osmotic up regulation of around 500 genes using the comprehensive proteomics and transcriptional profiling of *B. subtilis* wild type strain during salt shock. The current study also investigates the whole genome transcriptional profiling of *B. subtilis* using tiling array data of osmotic shock and high osmolar conditions, along with the comparison of these regulated genes to previous studies.

1.4 DNA MICROARRAY TECHNIQUES

Microarrays are the most commonly used techniques to measure the expression levels of large number of genes and often the whole genome. The microarray is a collection of orderly arranged microscopic DNA spots that are attached to a solid surface usually glass or nylon substrates. Based on the size of the spots, the array is termed as microarray or macroarray. Each DNA spot contains picomoles of a specific DNA sequence, called as probes/reporters. These probes can be small oligonucleotides,

cDNAs or DNA sequences. These probes are hybridized to a fluorescently labeled target sample either cRNA or cDNA. A scanner then detects the probe-target hybridization and then subsequently quantifies the signal. The data thus obtained is processed by the application of bioinformatics and data mining approaches. As the size of spots is small, a single array can accommodate a large number of several probes, thus facilitating the comparative analysis of transcriptome data in parallel, making the technology widespread.

In addition to the expression analysis, microarrays have several applications in various fields including gene discovery, disease diagnosis, pharmacogenomics, toxicogenomics etc. Thus, according to their application, microarrays are designed in several ways. One of the main features that play an importance role in the design is the fabrication. The probes are fabricated on the chip by photolithography using pre-made masks or dynamic micromirror devices, printing onto glass slides, pipette, drop-touch, piezo-electric (ink-jet), and electrochemistry on microelectrode masks. Custom-made arrays are also available from various manufacturers like Affymetrix, Agilent etc. The current study uses the data obtained from the spotted oligonucleotide as well as tiling arrays.

1.4.1 SPOTTED OLIGONUCLEOTIDE ARRAY

For spotted microarrays, oligonucleotides, cDNA or small fragments of PCR products corresponding to mRNAs are selected as probes. The oligonucleotides are short fragments of the single stranded DNA molecules. The size of the probes varies in different sizes, usually 60-mer probes are used in the Agilent design and 25-mer probes by Affymetrix platform. These probes are designed to match either known or predicted open reading frames. The main principle behind the spotted arrays is that the probes are synthesized prior to the deposition on the surface and are then spotted using fine-pointed pins or needles onto the glass. This printing step is usually performed using robotic arms to acquire high precision spotting of each probe at a designated location on the array surface. These arrays are then hybridized with corresponding cDNA or cRNA targets. Such spotted arrays are cost effective and mainly used in the cases where custom selected probes are used.

1.4.2 TILLING ARRAY

The tiling arrays are used to make out new insights into the transcriptome and methylome. These arrays are also based on the same principle as microarrays, the quantification of probe-target hybridization, in which the labeled targets are hybridized to the non-labeled probes attaches on a solid surface. The probe design is the main added advantage of the tiling arrays. The probes are designed in such a way that they cover the entire genome, with fixed / varying probe length and a particular distance (spacer) to the adjoining probes. Each array therefore contains up to millions of copies of single probe. Thus, the design of tiling array will allow global mapping of transcribed regions and will be useful in predicting new transcription segments that have not be seen before. This technology also enables mapping of both strands with a high resolution and sensitivity. *B. subtilis* tiling arrays were designed by Hanne Jarmer (Rasmussen *et al.*, 2009). These DNA tiling arrays consists of around 385,000 isothermal probes with a probe length of 18-22 nucleotides and a spacer of 22 nucleotides. The transcriptional analysis of the data is performed using a model of signal up-shifts and the drifts (Nicolas *et al.*, 2009).

1.5 MICROARRAY DATA ANALYSIS USING CLASSIFICATION ALGORITHMS

One of the major challenges in microarray data analysis is the selection of suitable data algorithms to analyze the huge amounts of generated data. The classical ANOVA is based on the assumption that the samples are in normal distribution and the random variables have equal variances. As these assumptions are not necessarily true for microarray data, classification algorithms are suggested to outperform these classical statistical analysis (Demsar, 2006) Classification algorithms can be separated into unsupervised learning methods such as K-means/K-medians and hierarchical clustering algorithms, and supervised learning methods like linear discriminant analysis, maximum likelihood discrimination, k-nearest neighbor methods, decision trees and random forests. Machine learning algorithms like support vector machines and random forests are nowadays widely used in high-level analysis of microarray data and data mining. The advantages of the random forest algorithm to classify huge amounts of data with a high number of attributes made its application in transcriptional analysis reasonable. Examples therefore are the prediction of the interactions between HIV-1 and human

proteins using gene expression data (Klein-Seetharaman *et al.*, 2009), the analysis of differential gene expression (Diaz-Uriarte & Alvarez de Andres, 2006), the analysis of mass spectrometry profiles (Barrett & Cairns, 2008; Datta, 2008; Izmirlian, 2004), the diagnosis of ulcerative colitis based on gene expression data, the detection of cancers, the prediction of child hood leukemia using gene expression data, and the prediction of protein-protein interactions (Chen & Liu, 2005). For the analysis of spotted oligo array data we implemented the random forest algorithm to classify genes that show a SigB-dependent expression in a wild type compared to its isogenic *sigB* knockout mutant.

1.5.1 RANDOM FOREST ALGORITHM

Leo Breiman and Adele Cutler first implemented the supervised machine-learning algorithm, Random Forest (RF) (http://www.stat.berkeley.edu/~breiman/Random Forests). RF is mainly based on the CART classification trees (Classification And Regression Trees). The CART trees use ensemble-learning methods like boosting, bagging etc. The successive trees gives extra weight to points incorrectly predicted by earlier predictors in boosting, whereas the successive trees does not have any influence/impact on the earlier trees in bagging. During the voting of the trees, a weighted vote is taken for the prediction in boosting, whereas a simple majority vote is taken for the prediction in bagging.

RF is mainly based on the bagging method, adding an additional layer of randomness to bagging (Breiman, 2001). The basic idea of RF is to construct trees using random vectors sampled from a data set. The trees were trained using a subset of randomly selected instances (genes; 63%) and variables, thus each tree is therefore considered a weak classifier. The forest of trees, however, is a strong classifier and ensures that an RF model can hardly be over trained. The random sampling used for training is quite similar to bootstrapping. The classification performance of a given tree is tested using a test set, the instances that were not used for training the tree. RF therefore determines the classification error during training which circumvents the use of cross validation tests, e.g., jack-knifing (Efron, 1982). An aggregate estimate of the error rate of the grown trees at each bootstrap iteration during the prediction of the data, which were not in the bootstrap sample, is termed as "out-of-bag" (OOB) error. The OOB error is

quite accurate and stable for the optimal number of trees grown in the forest (otherwise, it fluctuates biasing upward).

In addition to the classification, RF also gives the measure of importance of variables used in the classification. While constructing trees in the RF, at each node the attribute causing the highest decrease in the Gini index is chosen as split. Gini feature importance is derived from the Gini index and is the sum of all decreases in the forest due to a given feature, normalized by the number of trees in the forest (Breiman, 2001; Klein-Seetharaman *et al.*, 2009; Strobl *et al.*, 2007).

The construction of trees slightly changes with other methods, where each node is split using the best split among all variables. Thus, with this strategy, RF is robust against over fitting and it outperforms well compared to discriminant analysis, support vector machines and neural networks. In addition to these, RF is a user-friendly and reliable classification model. The random forest (RF) classifier package (version 4.5-35) implemented in the statistical programming language R (version 2.8.0) was used in the present study.

1.6 SEQUENCE ANALYSIS

Next to the gene expression analysis the identification of promoter sequences of the σ^B regulon is a key aspect to elucidate the structure of the general stress response in *B. subtilis*. According to the cellular requirement, the transcription of a gene is initiated depending on its specific sigma factor. The bacterial sigma factors binds to RNA polymerase and subsequently recognize a defined DNA sequence, known as promoter, located upstream of a gene. Usually the two defined sequences of a promoter are centered at about 10 base pairs and 35 base pairs respectively upstream of the defined start site of transcription. These sequences are often called as -10 box and -35 box or regions or elements. Using sequence alignment of known / experimentally verified promoter sequences of the regulon the consensus sequences of the -10 box and -35 box and the ideal spacer length between both motifs can be determined. In spite of having these certain defined features, bacterial promoters vary in strength and sequence. The genome-wide search of these consensus promoter sequences is often done by using profile hidden markov models or window based search algorithms. The high variation

among the sequences (deviations from consensus), distinct spacer lengths etc., has made bacterial promoter identification challenging.

The current study uses a window based consensus whole genome promoter search, using Motif Finder (Decodon GmbH, Greifswald, Germany). At the defined positions in the whole genome, usually 300 bp upstream of the coding DNA sequence (CDS), the consensus search is performed. Because of the inaccurate start site, the search was also performed up to 100 bp downstream of CDS. The selection of the number of mismatches to the known consensus sequence is based on the search stringency. The optimal selection of the mismatches ranges from two to four. The spacer length can be adjusted to the defined number of nucleotides, usually 12-15 bp for the SigB promoter.

1.7 AIMS OF THE STUDY

The general stress response of *B. subtilis* is one of the most noticeable changes in the gene expression network. Even the structure of SigB regulon has been studied earlier by three major transcriptome studies, but the obvious overlap of these studies remained quite low. Thus, in this thesis, the prime motive was to determine the structure of the general stress regulon. This SigB-dependent general stress response was comparatively analyzed in a wild type and its isogenic mutant strain using a targeted DNA microarray analysis. A machine learning algorithm, Random forest was used for the classification of the SigB regulon members with the data obtained from the spotted oligonucleotide array querying wide variety of stress and starvation conditions, incorporating previous knowledge. Genes that are subjected to the control of secondary regulators were also determined.

In the second part of this thesis, the defined structure of the SigB regulon was combined with the gene expression data of a tiling array analysis. This analysis led to the identification of new SigB targets, determined by the high resolution tiling arrays. Along with the SigB mediated general stress response, the tiling array data have also been used to analyze the osmo specific responses of *B. subtilis*.

2. METHODS

2.1 COMPREHENSIVE TARGETED SPOTTED OLIGO ARRAY

The experimental analysis was performed by Falk Hertwig (Hertwig, 2003) and the oxygen limitation experiments by Maren Depke. The current study deals with the data preparation and the random forest model design. The culture conditions and the array design were mentioned in brief, to understand the experimental setup. The whole analysis was published in (Nannapaneni *et al.*, 2012) and the complete dataset is available via the gene expression omnibus at NCBI (http://www.ncbi.nlm.nih.gov/geo/) with the accession number "GSE32895".

2.1.1 BACTERIAL STRAINS AND CULTURE CONDITIONS

The strains used in the study are *Bacillus subtilis* wild type strain 168 Marburg (*trpC2*) (Burkholder & Giles, 1947) and its isogenic mutant BSM29 (*trpC2 sigB:: spec*). BSM29 was constructed by insertion of a spectinomycin resistance cassette into *sigB* (Brigulla *et al.*, 2003). 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, oxidative stress); energy stress (O_2 limitation and glucose limitation) and low temperature growth. Sampling started in the exponential growth phase, at an OD_{540} of 0.45 to 0.55 for both strains (corresponding to time 0 min). The various stresses were applied to the cultures incubated at 37°C with shaking at 210 rpm: 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°C to 48°C (Helmann et al., 2001); oxidative stress – the culture was treated with hydrogen peroxide in a final sub-lethal concentration of 60 µM (Helmann et al., 2003). Samples were taken in regular time intervals of 5, 10, 15, 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 an OD_{540} of 0.1 and then the cultures were shifted to 15°C

(Brigulla *et al.*, 2003). Afterwards, the cultures were sampled at an OD_{540} of 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 OD_{540} of 0.3 from 210 rpm to 50 rpm. The sampling for both glucose and oxygen limitation was conducted at the time intervals of 15, 30, 45, 60, 90 min after the induction of the limitation. Each of the cultivations was performed in duplicate, except for butanol stress where only one cultivation was performed.

2.1.2 ARRAY DESIGN AND HYBRIDIZATION

Of the total 350 genes included in the array probe design, 250 genes were described as SigB-dependent by either one 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 on the array. The array also included 52 genes predicted to be SigB-dependent by other literature studies (Petersohn *et al.*, 1999b; Petersohn *et al.*, 2001), 10 additional genes and 36 control genes not known to be controlled by SigB. ORF sequences for the 350 selected genes were exported from Genbank and used for the probe design by Genescan Europe (Freiburg, Germany). 50 nt long isothermal probes 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 earlier (Waldmüller et al., 2002). cDNA was synthesized and labeled as described previously (Jürgen *et al.*, 2005). The array hybridization was described in Hertwig, 2003 and Nannapaneni *et al.*, 2012. The schematic representation of the conditions tested and the hybridization design was shown in Fig. 4.
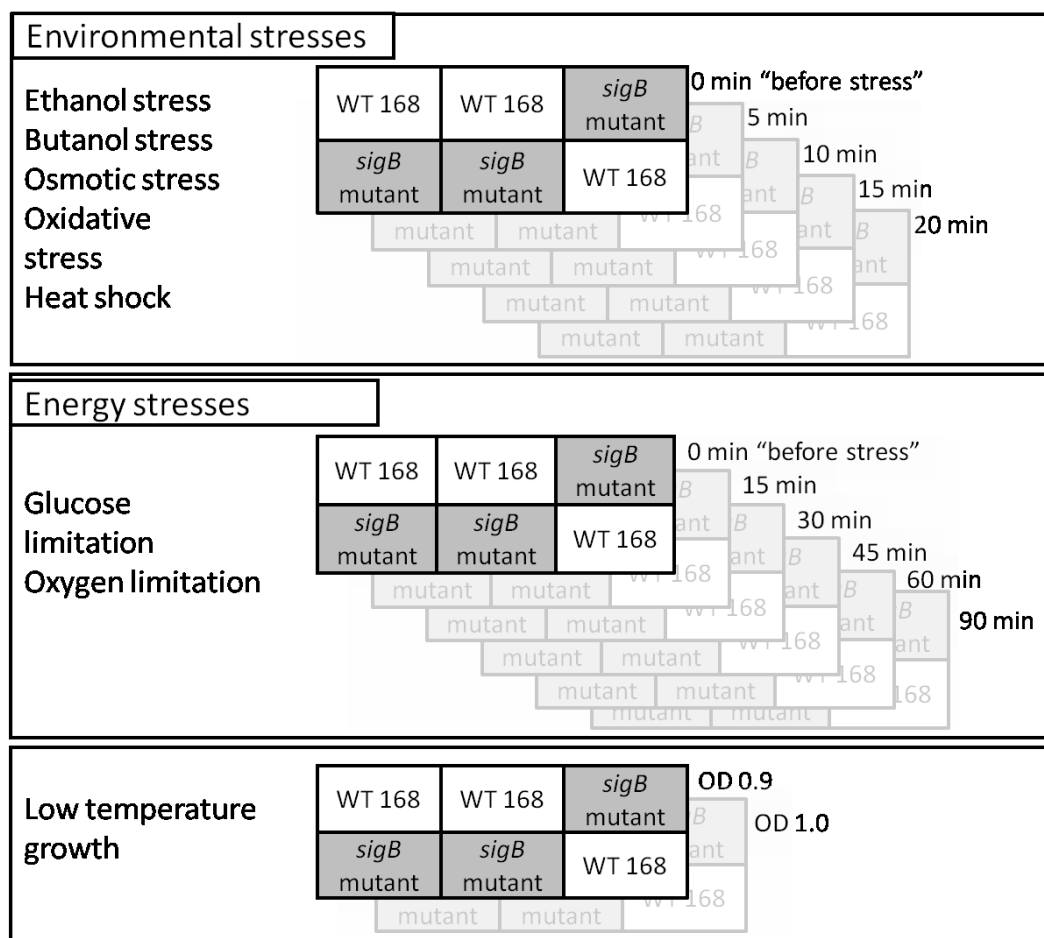
Fig. 4 Schematic representation of stress conditions and the time points tested in the study. Each stress was tested in three replicates, including one dye swap and three replicate spots. Each slide consists of wild type strain 168 and *sigB* knockout mutant BSM29 and the Cy5 and Cy3 labeling was represented as the white and grey colored boxes in the figure. Environmental stresses have five time points tested in the intervals of 5 min each, energy stresses have six time points in the intervals of 15 min and for the low temperature growth two samples at OD 0.9 and OD 1.0

2.1.3 RAW DATA ANALYSIS

After hybridization the slides were scanned using a ScanArray Express microarray scanner (Perkin Elmer Lifesciences, Boston, USA; 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 of the channels. From these images the signal intensities of individual spots were calculated using the Array Vision 6.0 program (GE Healthcare). The software (Array Vision 6.0) extracted the median trimmed density after subtraction of the 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 less than 2 or a difference of the median spot intensities between replicate spots greater than 1.5 fold were flagged as absent. Replicate spot intensities were then 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 median spot intensity values of the absent flagged spots were taken to obtain a complete data matrix for the random forest algorithm. Three replicates for each tested condition were analyzed, among them, one was the dye swap and the other two were biological replicates.

2.1.4 DATA PROCESSING

Generally, correction for systematic effects (e.g., dye effects) in dual dye microarray data is usually achieved using Lowess normalization. In the current hybridization design, the wild type and *sigB* knockout were always co-hybridized, consequently the gene expressions differs widely depending on stress condition and time point. The assumption of Lowess normalization that the majority of genes/probes has unchanged expression (50% in our study) is therefore not applicable for most of the hybridizations. Because of this, we used the synthetic slide method to normalize our data. Synthetic slides were created from samples having similar/equivalent gene expression pattern (over 50% of the probes have similar signals after normalization, e.g. Fig. 5). The equivalence of samples was determined using correlation analysis. Therefore, we separated the signals of the Cy3 and Cy5 channel (combined spot intensities) for all slides and did a correlation analysis using Pearsons product moment. Then, we paired these datasets into synthetic slides if their Pearsons correlation was above 0.8. To normalize for systematic effects, the paired datasets (synthetic slides) were normalized by Lowess (f = 0.5) using MicroPrep (van Hijum *et al.*, 2003). In the next step, the ratios of gene expression for the wild type strain with respect to the *sigB* knockout strain were calculated.
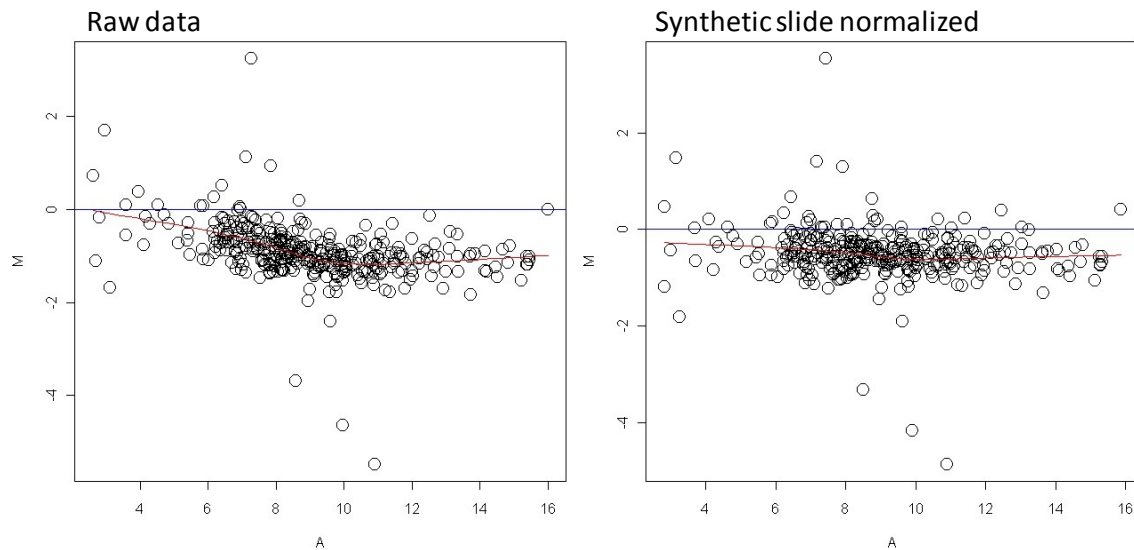
Fig. 5 MA plots (average intensity of the signal vs. log-ratio) of raw data and the corresponding synthetic slide normalized data. For this plot, 0 min time point of ethanol stress was used. After synthetic slide normalization, we expect the conditions (over 50% of the probes) should contain similar gene expression patterns.

2.1.5 RANDOM FOREST MODEL CONSTRUCTION

In the current work, two random forest models, the "expression RF" and the "kinetic RF" models were used for the identification of SigB regulon members. For both models, the overlap of three transcriptome profile studies (Helmann *et al.*, 2001; Petersohn *et al.*, 2001; Price *et al.*, 2001), 67 genes, and a set of 36 defined controls on the array were chosen as initial training set. In the "expression RF" model, gene expression ratios of wild type versus *sigB* knockout of a given stress and time point were used as variables. The "kinetic RF" model used gene expression ratios of the wild type strain 168 to its corresponding 0 min time point for each stress (exponential growth before applying the stress) as variables. The scripts that are used for the training and testing of the random forest models were shown in Appendix 2.1.5.

2.1.5.1 ANALYSIS OF VARIANCE (ANOVA) ON THE INITIAL TRAINING SET

In order to derive a stringent training of the models and to remove the number of false-positives in the prediction, both the models were subjected to an additional statistical analysis before the training. We chose to perform an ANOVA on the initial training set. For both the models, the initial training set of 67 SigB-dependent genes and 36 selected controls was subjected to ANOVA. This additional statistical analysis on the training set

filters based on the significant differences among the wild type and mutant strain in the "expression RF" model whereas it identifies significant differences among 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 a Welch t-test assuming that the variances were not equal. The p-value cutoff was 0.05 and the Benjamini and Hochberg false discovery rate was used for the multiple testing correction.

2.1.6 PROMOTER SEARCH AND SCORING

A window based consensus promoter search was conducted using Motif Finder (Decodon GmbH, Greifswald, Germany) for the identification of promoter sequences upstream of the predicted SigB-dependent genes. The motif search was performed 300 bp upstream and 100 bp downstream of the CDSs of the re-annotated *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.

The promoters of the putative SigB-dependent first genes in the operons (127) were aligned and a consensus sequence for the -10 motif and the -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 -10 motif and -35 motif. The resulting scores for all 127 genes were ranked in the ascending order and the ranking of the genes was 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 based on the highest expression ratio in the ascending order. In addition, for each gene, the mean rank of all these expression ranks for different stresses was calculated.

2.2 *BACILLUS SUBTILIS* WHOLE TRANSCRIPTOME ANALYSIS USING TILING ARRAYS

A systematic and quantitative investigation of the transcriptome of *B. subtilis* was performed within the framework of the EU-funded BaSysBio project and the BaCell-SysMO project. The transcriptome of *B. subtilis* exposed to 104 different growth and stress conditions was assessed by means of a high-density tiling array (BaSysBio *Bacillus subtilis* T1385K array version 1) designed by Hanne Jarmer (Rasmussen *et al.*, 2009). The study was published in (Nicolas *et al.*, 2012) and the complete dataset is available via the gene expression omnibus at NCBI (http://www.ncbi.nlm.nih.gov/geo/) with the accession numbers "GSE27219", "GSE27303", "GSE27419" and "GSE27652". The culture conditions are briefly summarized in the following section.

2.2.1 BACTERIAL STRAINS AND CULTURE CONDITIONS

All the experiments were performed using the BaSysBio reference strain BSB1 which is a tryptophan-prototrophic (Trp$^+$) derivative of the 168 *trpC2* strain (Anagnostopoulos & Spizizen, 1961). In total, 269 mRNA profiling experiments were performed under a large variety of conditions including growth conditions, stress conditions and carbon source shifts (Table 1). Growth conditions include: growth in rich medium (LB and LBG) and in minimal medium (M9), various carbon sources (M+G, Glu, Fru, Mal, G+S, Pyr, Gly and Glucon), sporulation induction (S0, S1, S2, S3, S4, S5, S6, S7, S8), germination (G135, G150, G180), competence induction (C30, C90), growth on agar plates (BC, BI), swarming cells (SW), biofilm formation (B36, B60, BT), anaerobic growth conditions (aero, nit and ferm), high cell density fermentation (T -0.4h to 5.0h), high and low phosphate growth (HPh, LPh, LPhT). Stress conditions include: continuous growth at high temperature/low temperature/high osmolarity (SMMPr, HiTm, LoTm, HiOs), heat/cold/ethanol stress (BMM, Heat, Cold, Etha), salt stress (Salt), mitomycin stress (M), diamide stress time course (dia) and oxidative stress (Diami, Paraq, H2O2, Oxctl). Carbon source shifts were from glucose to glucose+malate (GM) and from malate to malate+glucose (MG).

Table 1. The culture conditions for the samples used in the tiling array analysis.

	The experimental condition	The control condition
GROWTH CONDITIONS		
Standard growth in LB	The cells were grown in LB medium at 37°C under vigorous shaking in flasks. The samples were collected during the transient phase (OD_{600nm} around 1.4) and the stationary phase (OD_{600nm} around 2).	The control samples grown in LB medium were collecting during the exponential phase (OD_{600nm} around 0.5).
Standard growth in LB+Glucose	The cells were grown in LB medium supplemented with 0.3% glucose at 37°C under vigorous shaking in flasks. The samples were collected during the transient phase (OD_{600nm} around 1.4) and the stationary phase (OD_{600nm} around 2).	The control samples grown in LB medium with 0.3% glucose were collecting during the exponential phase (OD_{600nm} around 0.5).
Sporulation induction	The cells were grown in CH medium and sporulation was induced by resuspension in sporulation medium. The samples were collected in a time series after induction of sporulation (1h, 2h, 3h, 4h, 5h, 6h, 7h, 8h).	The samples were collected at the 0h time point before sporulation.
Germination	Germination was initiated by addition of 10 mM L-alanine to heat-activated spores. After 1 h, LB medium was added and samples were taken at different time points during the germination process (135 min, 150 min, and 180 min after addition of L-alanine).	Time point 135 min after addition of L-alanine
Competence induction	Grown on LB agar plates, the cells were inoculated in SPI medium at 37°C (OD_{600nm} of 0.5). At the end of exponential phase, the cells were diluted ten times in SPII medium. The sample was taken after 90 min, the maximum competence after the dilution.	The control sample was taken at the time point (30 min after the dilution) when no competence was detectable.
Growth on agar plates	Growth on LB agar plates, confluent colonies (around 10000 colonies per plates).	Cells from plates containing individual colonies (~ 100 colonies per plates).
Swarming cells	Cells from exponential growth were spotted on 1 % agar LB plates and incubated 16h at 37°C.	Cells from plates containing individual colonies (~ 100 colonies per plates).
Biofilm formation	Cells that were grown in 10 ml LB medium for 10h at 30°C were diluted in 10ml of MSgg medium and were incubated without shaking at 30°C. Biofilms were collected from 96 well plates after incubation for 36h and 60h.	Cells were incubated under vigorous shaking in MSgg medium and collected after 24h of incubation.
Standard growth in minimal medium (M9 medium)	Cells were grown in M9 supplemented with 0.3 % glucose under vigorous shaking. The samples were collected during the transient phase (OD_{600nm} around 1.5) and the stationary phase (OD_{600nm} around 1).	Samples were collected during exponential phase (OD_{600nm} around 0.5).

	The experimental condition	The control condition
Effect of various carbon Sources	Precultures in M9 medium were used to inoculate 100 mL M9 medium in 500 mL non-baffled shake flasks at a starting OD_{600nm} of 0.02. These are supplemented with - -L-Malic acid + D-glucose (3 and 2g/L) -D-Fructose (3g/L) -L-Malic acid (4.5g/L) -Glutamic acid + Succinic acid (2 and 2g/L) -Pyruvate (6g/L) -Glycerol (6g/L) -D-Gluconate (4g/L).	The cultures that are supplemented with -D-glucose (3g/L).
Anaerobic growth conditions	Cells were grown in synthetic medium in completely filled flasks with rubber stoppers and shaking at 100 rpm, in the presence (nitrate respiration) or absence (fermentative growth) of 10mM nitrate.	Cells grown aerobically in a 500 ml baffled flask with shaking at 250 rpm.
High cell density fermentation	During a batch fermentation with a non-sporulating *B. subtilis* strain in modified M9 medium (maximum OD_{600nm} of around 10). 7 samples at various times before exhaustion of glucose (-5.70h to -0.70h) and 9 samples at various times after glucose starvation (0.5h to 5.0h) were collected.	The time point when glucose was exhausted (0h).
High / low phosphate growth	Cells were grown in low phosphate medium and samples were collected 3h after the outset of phosphate-limitation induced stationary phase. Cells were grown in high phosphate medium and samples were collected during exponential phase (OD_{600nm} of 0.25).	Samples were collected during exponential growth in low phosphate medium.

STRESS CONDITIONS		
High temperature growth	Cells were grown in Spizizen's minimal medium (SMM) at 37 °C under vigorous agitation. Precultures were diluted to an OD_{578nm} of 0.1 and transferred to 51°C. Cells were harvested at OD_{578nm} of 1.	The control cultures were grown at 37 °C in SMM medium.
Low temperature growth	Cells were grown in Spizizen's minimal medium (SMM) at 37 °C under vigorous agitation. Precultures were diluted to an OD_{578nm} of 0.1 and transferred to 16°C. Cells were harvested at OD_{578nm} of 1.	The control cultures were grown at 37 °C in SMM medium.

	The experimental condition	The control condition
High osmolarity growth	Cells were grown in Spizizen's minimal medium (SMM) at 37 °C under vigorous agitation. Precultures were diluted to OD_{578nm} of 0.1 in SMM medium containing 1.2M NaCl. Cells were harvested at OD_{578nm} of 1.	The control cultures were grown at 37 °C in SMM medium.
Heat stress	Cells were grown in Belitsky minimal medium (BMM) at 37 °C under vigorous agitation. Heat stress was applied to exponentially growing cells at OD_{578nm} of 0.4. Samples were collected 10 min after a rapid temperature up-shift from 37 °C to 48 °C.	The control samples (BMM) were collected before the application of heat stress.
Cold stress	Cells were grown in Belitsky minimal medium (BMM) at 37 °C under vigorous agitation. Cold stress was applied to exponentially growing cells at OD_{578nm} of 0.4. Samples were collected 60 min after a temperature down-shift from 37 °C to 18 °C.	The control samples (BMM) were collected before the application of cold stress.
Ethanol stress	Cells were grown in Belitsky minimal medium (BMM) at 37 °C under vigorous agitation. Ethanol stress was applied to exponentially growing cells at OD_{578nm} of 0.4. Samples were collected 10 min after imposing ethanol stress by adding ethanol to a final concentration of 4% (v/v).	The control samples (BMM) were collected before the application of ethanol stress.
Salt stress	Cells were grown in Spizizen's minimal medium (SMM) at 37 °C under vigorous agitation. Salt stress was applied to exponentially growing cells with OD_{500nm} of 0.4. Samples were collected 10 min after imposing salt stress by adding NaCl to a final concentration of 0.4M.	The control samples (SMM) were collected before the application of salt stress.
Mitomycin stress	Cells were grown in LB medium at 37 °C under vigorous agitation. Mitomycin stress was applied to exponentially growing cells by adding 40ng/ml mitomycin. Samples were collected at 45, 90 min after the addition of mitomycin.	The control samples were collected before the application of mitomycin stress.
Diamide stress time course	Cells were grown in LB medium (OD_{600nm} around 0.6) at 37 °C under vigorous shaking. Samples were collected 5 min and 15 min after addition of diamide (0.5mM).	Cells harvested before diamide addition were used as unstressed control condition.
Oxidative stress	Cells were grown in LB medium at 37°C. At OD_{540nm} of 0.3, samples were taken 10 min after the addition of - Diamide 0.6 mM, - Paraquat 0.4 mM, - H_2O_2 0.1mM.	For the control culture no oxidative drug was added to the medium.

	The experimental condition	The control condition
CARBON SOURCE TIME SHIFTS SERIES		
Glucose to Glucose+Malate	The M9 medium precultures were inoculated in the M9 medium in a 3.1 L bioreactor. In the bioreactor, temperature was controlled at 37 °C, the pH was controlled at 7.2 and the dissolved oxygen tension was maintained above 50%. The cells were grown on the single substrate glucose up to an OD_{600nm} of 0.50, and then the second substrate malate (4 g/L) was added. The samples were taken during the nutrient shift from glucose to glucose+malate, (5, 10, 15, 25, 45, 60, 90, 120, 150 min after the addition of malate).	The samples were taken 0 min before the addition of malate.
Malate to Malate+Glucose	The M9 medium precultures were inoculated in the M9 medium in a 3.1 L bioreactor. In the bioreactor, temperature was controlled at 37 °C, the pH was controlled at 7.2 and the dissolved oxygen tension was maintained above 50%. The cells were grown on the single substrate malate up to an OD_{600nm} of 0.50, and then the second substrate glucose (3g/L) was added. The samples were taken during the nutrient shifts from malate to malate+glucose, (5, 10, 15, 25, 45, 60, 90, 120, 150 min after the addition of malate).	The samples were taken 0 min before the addition of glucose.

2.2.2 TILING ARRAY RAW DATA ANALYSIS

The transcription signal associated with each probe on the array was calculated using a probabilistic hidden markov model, which calculates based on signal shifts and drifts (Nicolas *et al.*, 2009). The intensity values for annotated genes were derived from the individual probe data by calculating the median of the probes located within the genomic coordinates of the genes. The newly identified features which includes the transcribed regions that are located outside the annotated CDSs as well as the RNA genes are detected when the locally reconstructed signal exceeds 10x the median of all probes (chromosome median), in at least one hybridization. Our colleagues in France, especially Pierre Nicolas, processed the raw data; a detailed description of the data processing is provided in Nicolas *et al.* (2012) and a general description is also given at the beginning of chapter 3.2 of this thesis.

2.2.3 TILING ARRAY EXPRESSION DATA ANALYSIS

The normalization of the chromosome expression profiles was selected to avoid interpreting experiment specific variations. Additionally, the condensed, linear space gene expression values were imported via the Gene Expression Text Loader (GETL) into the Rosetta Resolver software (Rosetta Biosoftware, Seattle, WA, USA) for data analysis. Inter-chip median-scaling (normalization) and detrending (Rosetta Resolver proprietary algorithm) were applied in the experiment definitions (ED) to allow direct comparison of induction/repression ratios with previous microarray platforms.

3. RESULTS

3.1. DEFINING THE STRUCTURE OF THE GENERAL STRESS REGULON OF *BACILLUS SUBTILIS* USING TARGETED MICROARRAY ANALYSIS AND RANDOM FOREST CLASSIFICATION

Gene expression was queried for most of the conditions described to be relevant with respect to SigB activation. These are ethanol-, butanol-, osmotic- and oxidative stress, heat shock, low temperature growth, glucose as well as oxygen limitation. To attribute the effect of SigB on gene expression, the *B. subtilis* wild type strain 168 and its *sigB* knockout mutant strain BSM29 were comparatively analyzed under all these conditions in short time series. The probes on the custom array targeted a total of 350 genes.

The selection of genes was based on:

(i)   67 genes that were described to be SigB-dependent in all the three transcriptional profiling studies (Helmann *et al.*, 2001; Petersohn *et al.*, 2001; Price *et al.*, 2001),

(ii)   185 genes that were reported to be SigB-dependent in either one or two of these three studies,

(iii) 52 genes reported in other studies to be probably SigB-dependent (Petersohn *et al.*, 1999b; Petersohn *et al.*, 2001; Price *et al.*, 2001)

(iv)   10 additional genes of interest based on unpublished data and

(v)   36 genes that were selected as controls not exhibiting SigB-dependent expression but coding for a variety of other cellular stress adaptation reactions (Appendix 3.1.1).

For all the stress conditions applied, the expression data for both the wild type and the mutant 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 min, 10 min, 15 min and 20 min) after imposing the stress were analyzed. The time points for glucose limitation and oxygen limitation were immediately before and 15 min, 30 min, 45 min, 60 min or 90 min after limitation of the nutrients. For the assessment of the gene expression pattern during low temperature growth at 15°C, samples were taken during mid-exponential growth at

OD$_{540}$ 0.9 and 1.0. In the design of the study we specifically re-used stress imposition schemes that had already been reported before (Antelmann *et al.*, 1996; Helmann *et al.*, 2003; Völker *et al.*, 1995). The stimuli for the environmental stresses were adjusted to intermediate strengths that do not completely prevent growth.

To classify "SigB" regulon members and "non-SigB" regulon members we used the random forest algorithm, a decision tree-based machine-learning algorithm. The model was trained using a training set derived from previous knowledge derived in the three-transcriptome studies (Helmann *et al.*, 2001; Petersohn *et al.*, 2001; Price *et al.*, 2001). This training set enabled us to combine the microarray data of our study with prior knowledge. The training set comprises two groups, "SigB" and "non-SigB". 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 of Price, Petersohn, and Helmann (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 were selected as "non-SigB" group. The remaining genes were placed in the test set, and the trained random forest model was then used to classify genes in the test set into either "SigB" or "non-SigB" based on the expression profiles.

The SigB regulon members were identified using two random forest classification models. The first one was based on the comparison of expression patterns in the wild type and the *sigB* mutant and 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 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 designated to screen for induction by a wide variety of stress stimuli to classify SigB from non-SigB genes.

3.1.1 DEFINING THE SIGB REGULON USING THE "EXPRESSION RF" MODEL

The training set in the "expression RF" model consists of two groups. The ANOVA analysis on the training set filters based on the significant differences among the wild type and mutant strain. The initial time point before applying stress (0 min time point) for wild type strain and mutant strain was excluded in this analysis as there will be no induction at this time point. The analysis thereby resulted in identifying 53 genes (out of the 67 genes recognized by all three previous transcriptional profiling studies) having significant differences among the wild type and mutant strain and 33 control genes having no significant differences.

The ANOVA filtered training set, 53 genes named as "SigB" group and 33 control genes named as "non-sigB" group. The fourteen genes (*csbC, csbX, dps, mcsA, mcsB, ybyB, ycdF. ycdG, ydbD, ygxB, yjbD, yqhB, yvyD, yxlJ*) were removed from the "SigB group" of the training set by ANOVA and the three genes (*spoIIQ*, *spoIIR*, *hom*) were removed from the "non-SigB" group. Taking the gene expression ratios between wild type and *sigB* mutant of the different stress conditions and time points as the 111 variables, the random forest classification model was trained based on the training set (53 "SigB" genes + 33 "non-SigB" genes). The out-of-bag (OOB) error of the training was reported as 1.16% indicating that only a little over 1% of the genes in the training set were incorrectly classified based on the microarray data. The subsequent testing of the model with 247 genes of the test set resulted in the prediction of 99 genes to the "SigB" group and 148 genes to the "non-SigB" group (Fig. 6A).
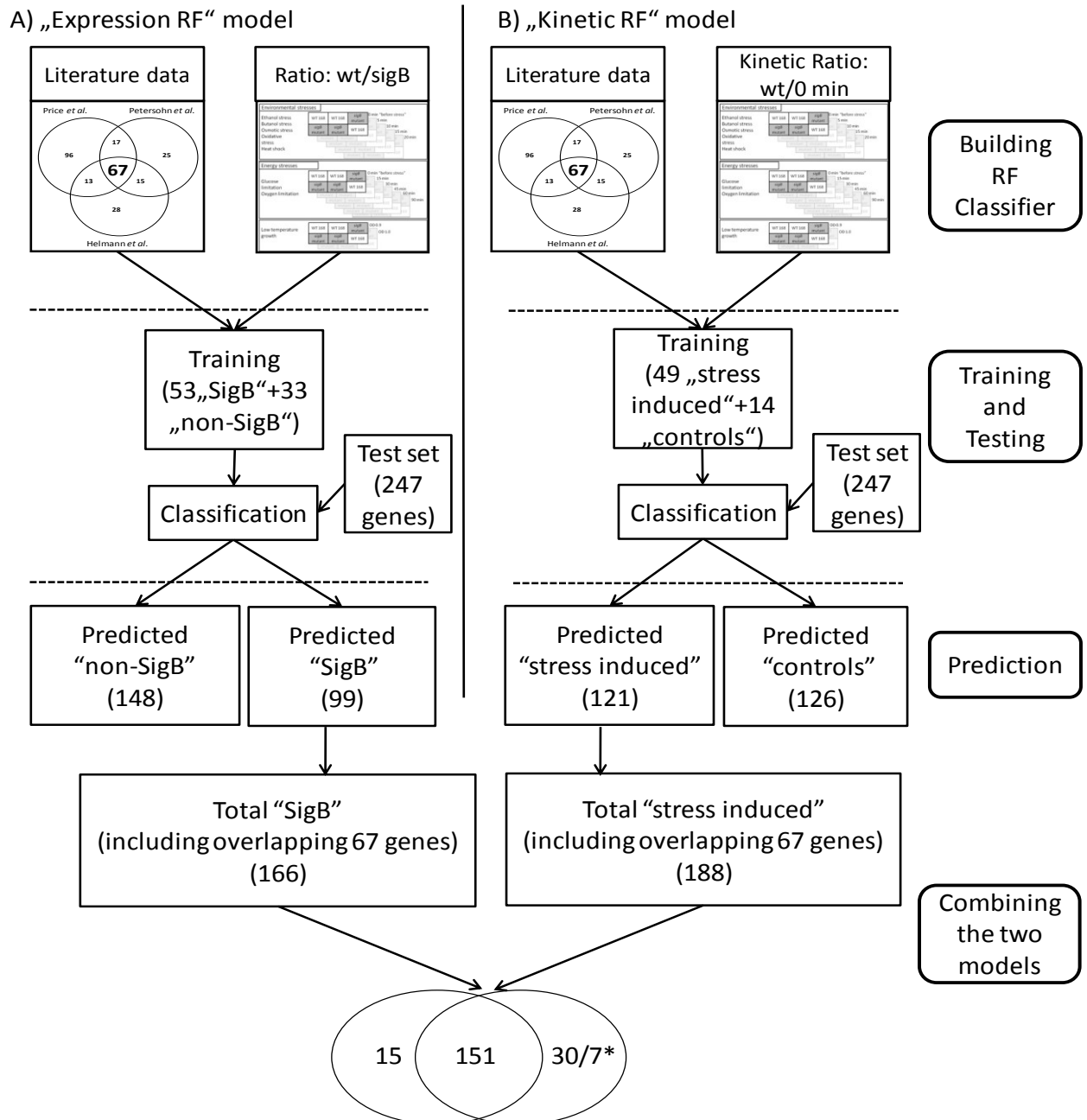
Fig. 6 Schematic representation of the random forest analysis. The random forest (RF) classifier was built based on literature data and array data. Both models were trained and used to predict the (non) SigB members. (A) "expression RF" analysis where the ratios of gene expression of wild type versus *sigB* knockout for all the stresses were taken as variables. (B) "kinetic RF" model where the variables were ratio of gene expression of 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 both models were compared after including the overlapping genes (67 genes) to each group.

Based on the class probability in the random forest, the genes in the test set were assigned to one of the two groups. In general, when the class probability of a gene is greater than 0.5, the gene is assigned to the class. For example, the class probability of the test set was plotted against the expression ratios for ethanol stress condition (Fig. 7). This plot indicates that there is a good correlation between the predicted class probabilities to the expression ratios. In addition, a more stringent classification resulted in higher probabilities. In the current study, the general class probability of >0.5 was taken into account. Therefore, applying the "expression RF" model, classification resulted in 166 SigB-dependent genes, 67 genes already unequivocally identified by the three transcriptional studies plus 99 genes predicted in this study (Appendix 3.1.2).
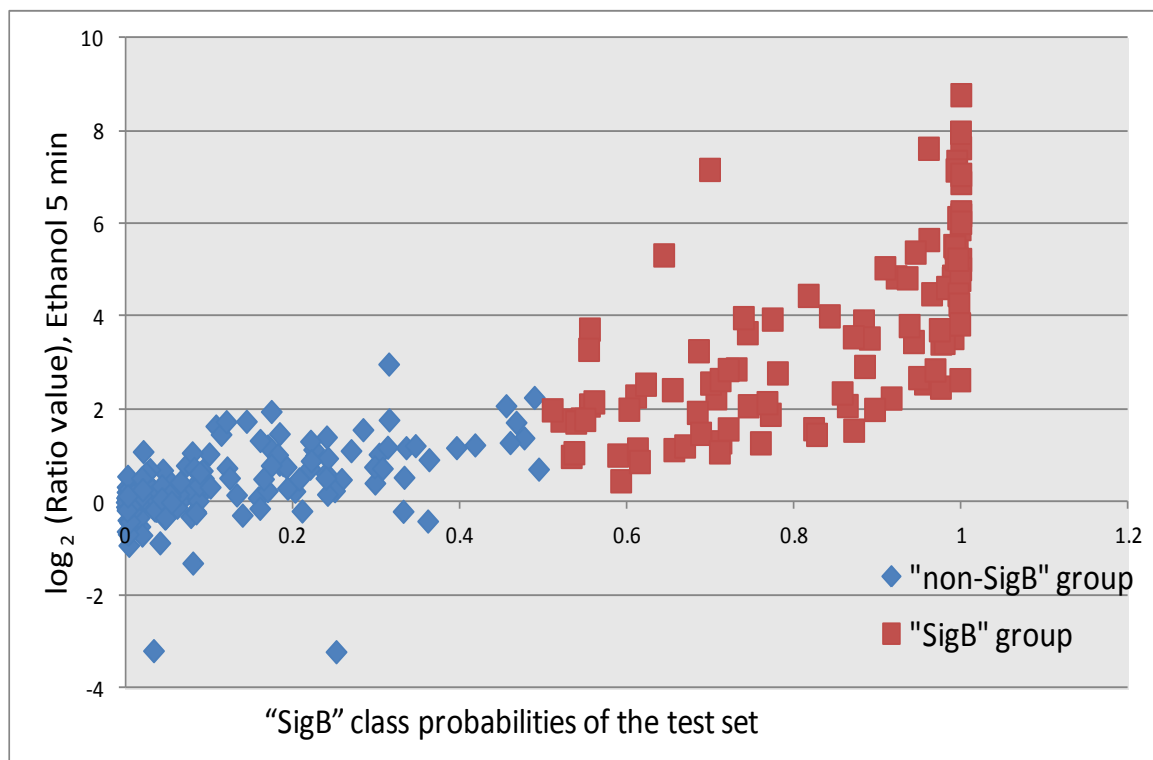
Fig. 7 Scatter plot representing the class probabilities of "SigB" vs. expression ratios. The blue colored triangles represents "non-SigB" group of the test set and red colored squares represent the "SigB" group of the test set. The wild type to mutant expression ratios for the ethanol stress condition at 5 min time point were represented in log 2 scale.

In addition to the prediction, the random forest algorithm provides information about the importance of the variables used (conditions) in the classification model. Based on the expression of the genes comprising the training set, the impact in the final classification for each stress condition was determined. This allows a ranking of the

stress conditions according to their potential to activate SigB (Fig. 8). For the stresses applied, in most cases the first time point after the implementation of stress was most important for 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 well with the importance of the time points to classify SigB dependence. Besides the kinetics of the SigB induction, also a hierarchy of the analyzed stress conditions can be discerned. The importance of butanol stress, osmotic stress and heat shock was higher than the importance of the other stresses, indicating that in general, SigB-dependent genes responded strongest during application of butanol, osmotic and heat stresses (Fig. 8). 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 compared to the variable importance calculated for RsbP-dependent energy stresses (Fig. 8).
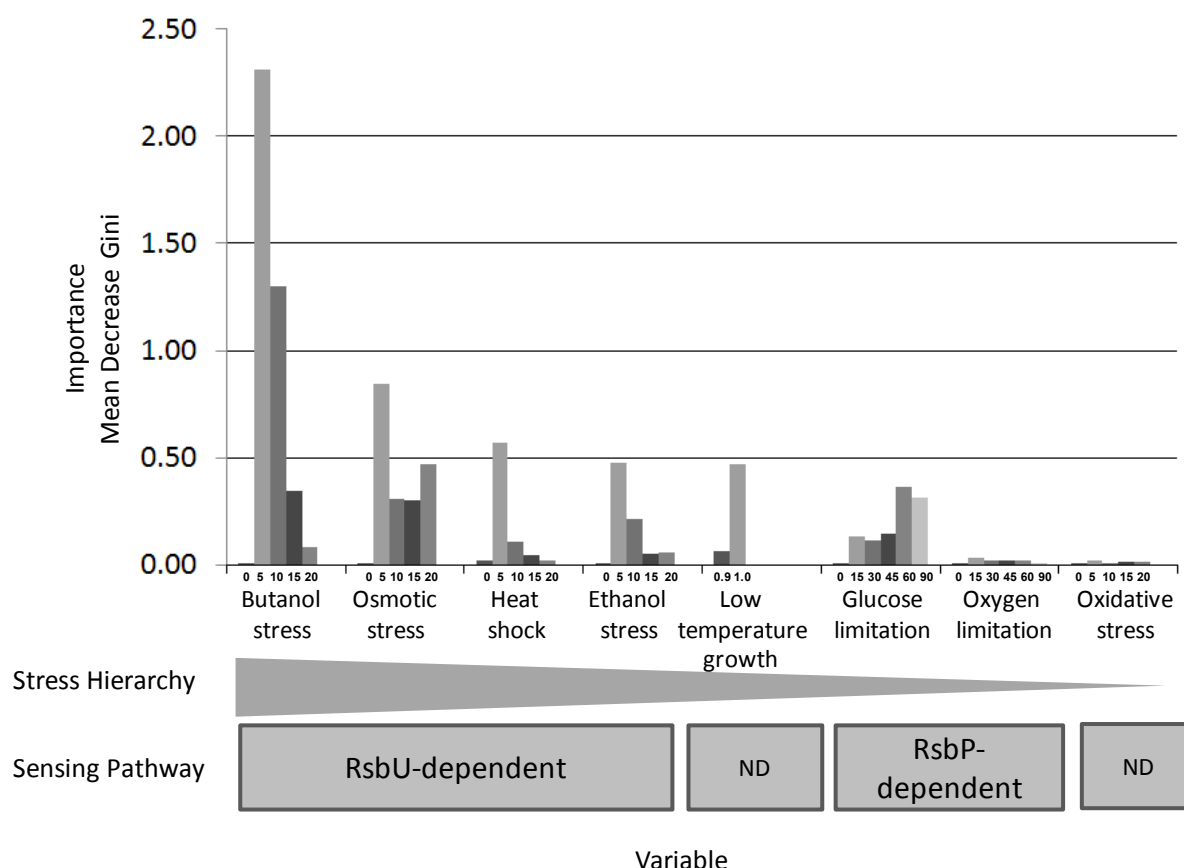
Fig. 8 Variable importance plot resulting from training the "expression RF" model. In this random forest model, variables are the expression ratios of wild type to *sigB* mutant strain for all the stresses. The bars in the plot represent the mean decrease Gini, a random forest generated variable importance measure. The higher the value of mean decrease Gini, the more important was the variable importance to differentiate between "SigB" and "non-SigB" group. The stress hierarchy, in which the order of stress conditions was plotted based on the maximum variable importance observed. When this importance is related to sensing pathways, the variable importance of RsbU-dependent environmental stresses was higher than the variable importance of RsbP-dependent energy stresses. Butanol stress has not been specifically assigned to any sensing pathway, with a high similarity to ethanol stress this was grouped along with RsbU-dependent stresses in the figure. The activation of SigB during low temperature growth does not depend on either RsbU or RsbP sensing pathway (Brigulla *et al.,* 2003). The signaling is also RsbV independent, so labeled as not defined (ND) in the figure. The sensing pathway involved during oxidative stress is also not defined (ND).

In addition to the importance of the variables of the training set, we plotted the expression ratios between wild type and *sigB* mutant of all 152 predicted and confirmed SigB-dependent genes for the time points displaying maximal induction. The expression-based ranking for ethanol stress was slightly higher compared to the variable importance based hierarchy. For all the other stresses, this expression ratio based analysis generated a very similar hierarchy of the stress conditions tested (Fig. 9).
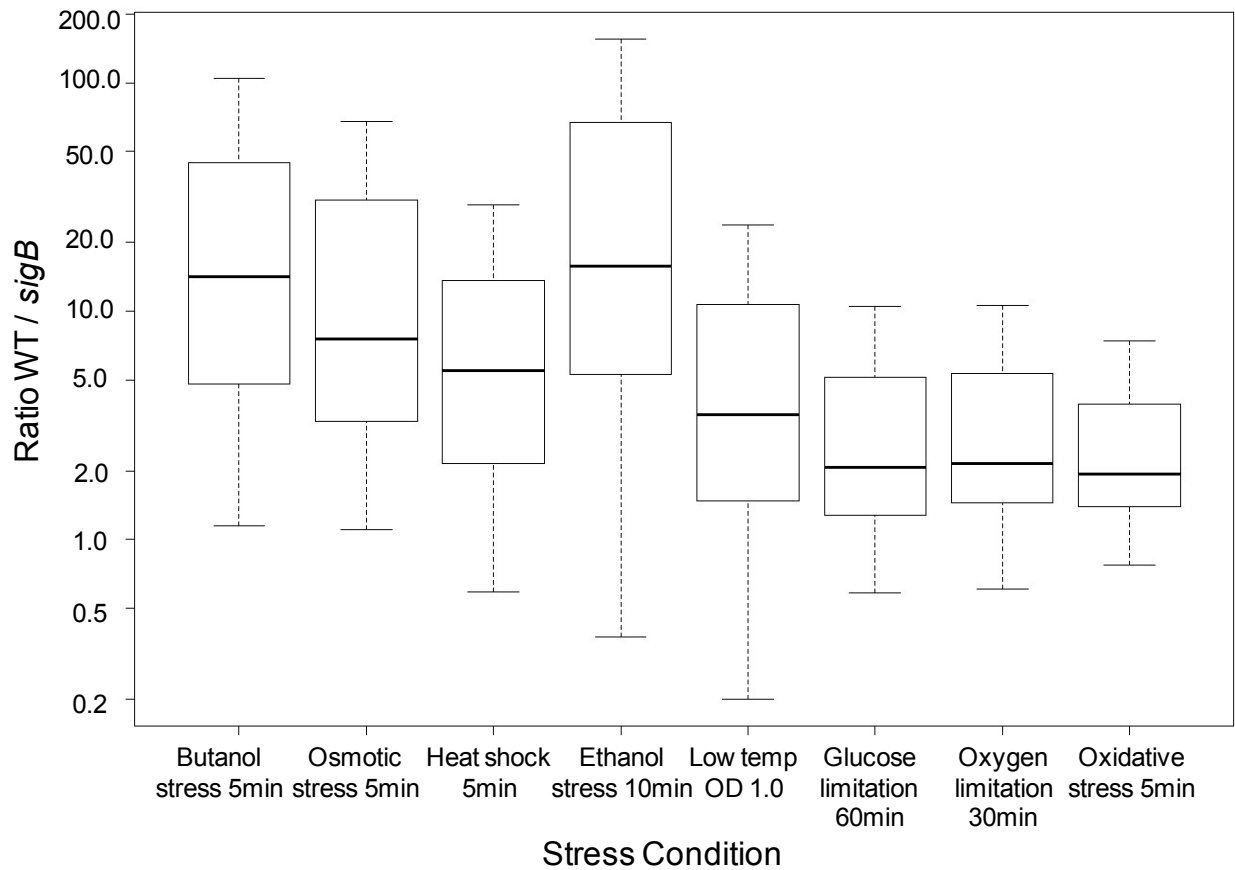
Fig. 9 Box plot of the expression ratios of wild type to mutant of the predicted SigB regulon members ("SigB" group from "expression RF model"). For a given stress condition, the time point where maximum induction was observed was plotted. The stress hierarchy observed in this plot with the expression profiles of the predicted SigB-dependent genes (152) was compared to the hierarchy from Fig. 8. The importance of ethanol was slightly higher in the expression box plot compared to the variable importance plot, all the other stresses were the same. The maximum induction of genes in ethanol stress was observed at time point 10 min in box plot, whereas the importance of 5 min time point of ethanol stress is still higher than 10 min time point in variable importance plot (Fig. 8).

3.1.2 IDENTIFICATION OF SIGB MEMBERS WITH POTENTIAL SECONDARY REGULATION USING THE "KINETIC RF" MODEL

The SigB-dependent genes (166 genes) that have clear expression differences among the wild type strain and its corresponding mutant strain were identified in the "expression RF" model. However, it is known that a part of SigB-dependent genes is controlled by secondary regulators as well (Krüger *et al.*, 1996; Petersohn *et al.*, 1999c; Scharf *et al.*, 1998; Varon *et al.*, 1996). These genes are induced under general stress conditions, even in the absence of SigB. When only focusing on expression differences between the wild type and the *sigB* mutant, SigB dependency of such genes would likely be missed. Therefore, a second RF model, the "kinetic RF" model, was trained with the wild type data to identify these additional regulon members for whom stress induced expression might also be seen in the *sigB* mutant strain. The training instances were carefully selected to ensure coverage and inclusion of the broad activation spectrum of SigB. Additionally, the stringent training of the model was required, and after classification, the selection of true SigB-dependent genes was of utmost importance. In order to achieve this SigB specificity of the candidate genes a putative SigB promoter, either known from 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. The ANOVA analysis identified 49 genes of the "stress induced" group as displaying significantly different expression after stress imposition compared to exponential growth and 14 genes of the "control" group where such significant differences were not observed. Usually, the expression differences between exponential growth and stress conditions were likely triggered by specific stress regulation, e.g. by osmotic stress for *opuAA* and *opuCA* in the control group. Such genes were removed from the control group from the pre-filtration by ANOVA. The "kinetic RF" model was thus trained with 49 genes as "stress induced" group and a set of 14 control genes as "control" group. The gene expression ratios of each stress condition to its corresponding 0 min time point in the wild type strain were taken as variables. The OOB error of training the "kinetic RF" was 12.7%. This OOB error of the "kinetic RF" is higher than that of the "expression RF" (1.16%). However, an OOB error of 12.7 percent is still

acceptable, indicating that the trained random forest model still can classify the two groups apart.

This trained model was then used to classify the test set of 247 genes based on the expression pattern in the wild type strain. A total of 121 genes were predicted as "stress induced" and 126 genes were predicted to be part of the "control" group (Fig. 6B). 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.

3.1.3 INTEGRATING THE PREDICTIONS FROM "EXPRESSION RF" AND "KINETIC RF" MODELS

The comparison of the classification predictions of "expression RF" and the "kinetic RF" models revealed a large overlap of 151 genes (Fig. 6). 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. 6). These genes still showed a considerable expression also in the *sigB* mutant strain and were therefore not detected as SigB-dependent in the "expression RF" model, indicating control by additional regulators. For 19 of the 37 genes, additional regulation by secondary regulators has already been proven by previous studies.

These 19 genes are: *bioB, bmrR, purK, radA, yacK, yacM, ydaF, ykrT, yocL, ypuD, yraA, ysnF, ywhH, yxjG* and the operon *yceC-yceD-yceE-yceF-yceG-yceH* (Au *et al.*, 2005; Banse *et al.*, 2008; Baranova *et al.*, 1999; Bower *et al.*, 1996; Chu *et al.*, 2006; Chumsakul *et al.*, 2011; Comella & Grossman, 2005; Derre *et al.*, 1999; Drzewiecki *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.*, 2010; 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 Appendix 3.1.3. The remaining 18 genes *nap, opuBA, opuBB, ydbB, ydjL, yfmG, yhaS, yhaT, yhaU, yisP, yqgE, yqiY, yqxH, ytzE, yvaC, ywaF, ywmE* identified in our study are also potentially controlled by a

secondary regulator. This assumption has been validated in the study of Nicolas et al., 2012.

As these 37 candidate genes were predicted solely because of their stress induction in the wild type, further evidence of their SigB dependency was required. We required the presence of a SigB promoter motif either identified in a window based DNA motif search or reported previously (Helmann *et al.*, 2001; Petersohn *et al.*, 2001; Sierro *et al.*, 2008). Since seven (*nap, ykrT, yocL, yqgE, ytzE, ywaF, ywhH*) out of the 37 genes failed to display such a SigB motif, the remaining 30 genes were considered to indeed possess SigB-dependent regulation (Appendix 3.1.3). Overall, including the remaining 30 genes, 196 SigB-dependent genes were identified in our study (Fig. 6). 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 (Appendix 3.1.4).

In order to compare the promoter strength to 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. 10A). As the ranks for each gene behaved quite similar in 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. 10B). The clustering resulted in four clusters (C1-C4). The two major clusters C1 and C4 harbor genes having good promoter scores (promoter similar to the consensus) and high expression ratios, or genes having low promoter scores and low expression ratios, respectively. The clusters C2 and C3 represent the low promoters scores with high expression ratios and high promoter scores with low expression ratios respectively. A detailed list of the clusters is provided in Appendix 3.1.5.
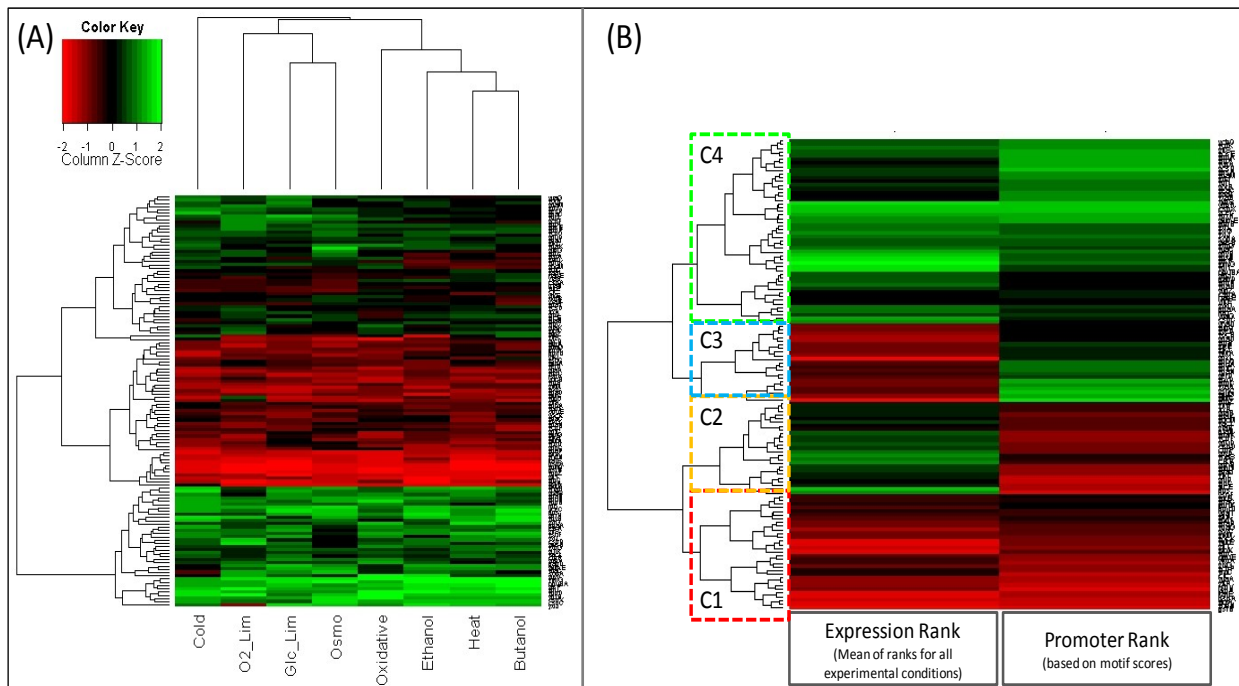
Fig. 10 Hierarchical clustering of genes based on the expression ratio ranks of the individual stress conditions after Z-scale transformation (A). Hierarchical cluster with indicated cluster C1-C4 of the mean of all expression ratio ranks and the corresponding promoter scores (B). For both analyses, 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 the ascending order as described in the methods section. Higher expression ratios, as well as the good promoter scores got the least ranks and thereby obtained the lowest Z-scores (colored in red). Lower expression ratios and low promoter scores had higher ranks and received higher Z-scores (colored in green).

Thus, the 196 SigB-dependent genes identified in the study were compared with the existing three major transcriptional profiling studies (Helmann *et al.*, 2001; Petersohn *et al.*, 2001; Price *et al.*, 2001). Around 95% of genes, which were described to be 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 our present study (Fig. 11). However, for genes reported to be SigB-dependent in only one of the three studies the frequency of validation by our current study was clearly lower.
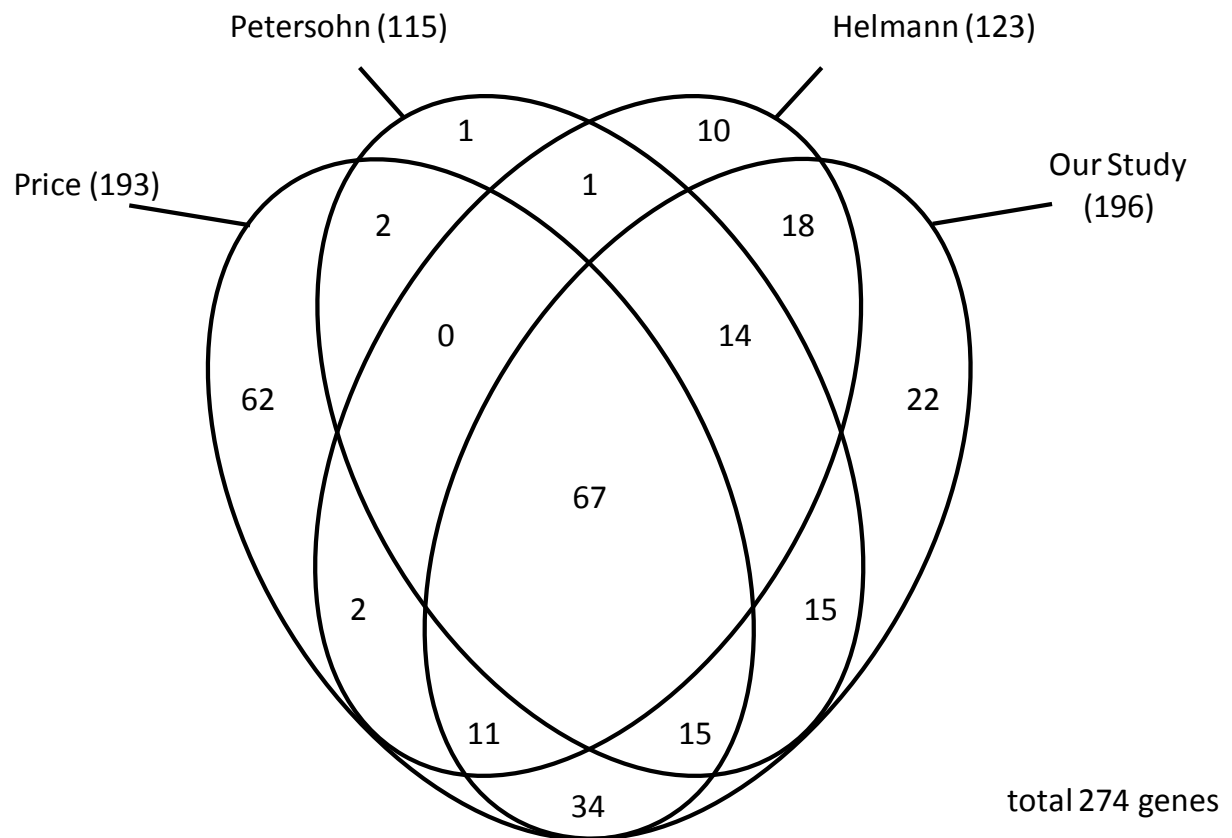
Fig. 11 Venn Diagram of the SigB gene lists from the previous three transcriptome studies and the current study. Around 95% of the genes described by at least two of the three earlier transcriptome studies were validated in our study.

3.2 BACILLUS SUBTILIS WHOLE TRANSCRIPTOME ANALYSIS USING TILING ARRAYS

Within two joint European projects, BaSysBio and BaCell-SysMO, an extensive transcriptome study was performed, in which a tryptophan-prototrophic (Trp$^+$) derivative of the *B. subtilis* 168 strain was analyzed under a wide range of nutritional and environmental conditions. The 104 different experimental conditions were selected to provide a broad coverage of the bacterium's lifestyles and include, for example, the utilization of various nutrients, anaerobic growth, biofilm formation, adaptation to various stresses, and conditions stimulating sporulation or the development of competence for genetic transformation. The contributions of our group included the analysis of *B. subtilis* during growth in M9 medium (exponential and stationary phases), under several stress conditions (high and low temperature, high salt concentrations, and ethanol) and during high cell density fermentation.

In this study, the whole transcriptome of *B. subtilis* was assessed by means of a high resolution tiling array. Altogether, 269 RNA samples were analyzed on a custom-designed tiling array (BaSysBio *Bacillus subtilis* T1 385K array) with a resolution of 22 bp (Rasmussen *et al.*, 2009).

Bioinformatic analysis of the 269 tiling array profiles was performed by our colleagues from INRA, Mathématique Informatique et Génome in Jouy-en-Josas (France), in particular Pierre Nicolas, and is described in detail in Nicolas *et al.* (2012). In brief, the transcription signal associated with each probe on the array was calculated from the raw data using a statistical model describing the transcription profiles in terms of abrupt shifts and progressive drifts (Nicolas *et al.*, 2009). For the detection of transcribed regions outside the annotated CDSs and RNA genes, the locally reconstructed signal had to exceed 10x the median of all probes (chromosome median), in at least one hybridization. In parallel, a repertoire of probable 5'- and 3'- transcript ends was established across the 269 profiles. Up-shift positions were defined as the coordinates of the 5'-end of the first probe at which the transcription signal increases. When up-shifts at adjacent probes were detected in different hybridizations, they were considered as the same transcript start, and the consensus position for the up-shift was computed as the average position. Transcribed regions were then decomposed into

segments according to the genome annotation (Barbe *et al.*, 2009) and to the detected transcript ends. By automatic and manual curation, a final list of 1583 previously un-annotated RNAs (S1 to S1583) was obtained, and these RNAs were classified according to their structural relationships with the neighboring annotated CDSs. The categories include, for example, 5'-UTRs, 3'-UTRs and independent segments defined as transcribed from their own promoter and ending with a down-shift. New RNA features overlapping annotated genes on the opposite strand were classified as antisense RNAs when the overlapping region was either ≥ 100 bp or, for regions shorter than 200 bp, when it accounted for ≥ 50% of the length of the new RNA feature or the annotated feature on the opposite strand. By considering all independent segments, antisense RNAs, and predicted CDSs, the study identified 512 potential new genes. For each annotated CDS and newly defined transcribed region, an aggregated expression value was computed as the median signal intensity of probes lying entirely within the respective region and having a unique match on the genome.

The identified signal up-shifts revealed 3242 putative promoter sites. Among these 3242 up-shifts, 961 up-shifts were validated by independent experimental evidence based on i) the sigma factor binding sites reported in DBTBS (Sierro *et al.*, 2008) and ii) the ~600 transcription start sites mapped by a differential RNA-seq approach that discriminates primary from processed 5'-ends (Irnov *et al.*, 2010). Pairwise correlations between promoter activities estimated from the mRNA amount downstream of each up-shift were summarized in a promoter correlation tree generated by hierarchical clustering. Thus, the promoter tree was built independently of DNA sequence information. A new unsupervised algorithm was applied to model the bipartite motifs (sigma factor binding sites) by combining information from the DNA sequences around the upshifts (-60 bp upstream to 40 bp downstream) and from the promoter correlation tree (promoter activities). The obtained motif types were assigned to sigma factors based on the sigma factor binding sites reported in DBTBS. In this way, 2935 of the 3242 identified promoters were assigned to the corresponding sigma factors.

3.2.1 SIGB-DEPENDENT GENES AND SIGB-INDUCING CONDITIONS IDENTIFIED BY THE TILING ARRAY APPROACH

This genome-wide classification of promoters assigned 170 transcription units, comprising 282 annotated genes and 133 previously un-annotated RNA features, to SigB-dependent regulation, shown in Fig. 12. Among these 133 un-annotated RNA features, 50 were identified as antisense RNAs, which can potentially have an impact on transcription, translation and/or stability of their cognate sense mRNAs. The 133 previously un-annotated RNA features with their start and end positions on the genome, the list of conditions in which these features are highly expressed, and their antisense targets (if any) are listed in Appendix 3.2.1.
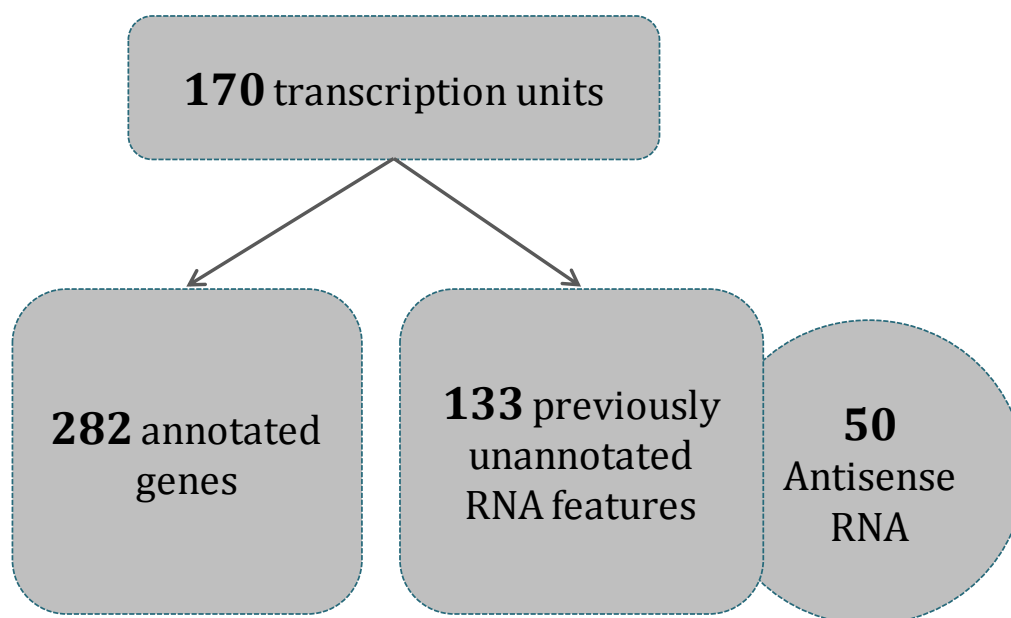
Fig. 12 Diagram showing the SigB-dependent features from the tiling array study. A total of 170 transcription units include 282 annotated genes and 133 previously unannotated RNA features i.e. the newly identified RNA transcripts. Among the 133 newly identified features, 50 of them were classified as the antisense RNAs.

When comparing the structural relationships of the 133 newly identified features with the annotated genes, 48 features represent 5′-UTRs (5′) and 19 features represent 3′-UTRs, comprising three subgroups: five features are 3′-untranslated regions (3'UTR), one feature is a 3′-extended transcript with no termination (3'NT) and 13 features are 3′-extended transcripts with partial termination (3'PT). Eight and 14 features, respectively, are RNAs independent of previously annotated genes, either displaying (indep) or lacking a termination site (indep-NT). In addition, 32 mRNA segments are

lying between two genes with distinct promoters (inter) and 11 are part of a polycistronic mRNA from a single promoter (intra). The distribution of all these newly identified RNA features is shown in Fig. 13A. As most of the newly identified features in the tiling array study are present along with the annotated genes in the same transcription units, they exhibit comparable induction levels as the annotated genes (Fig. 13B).
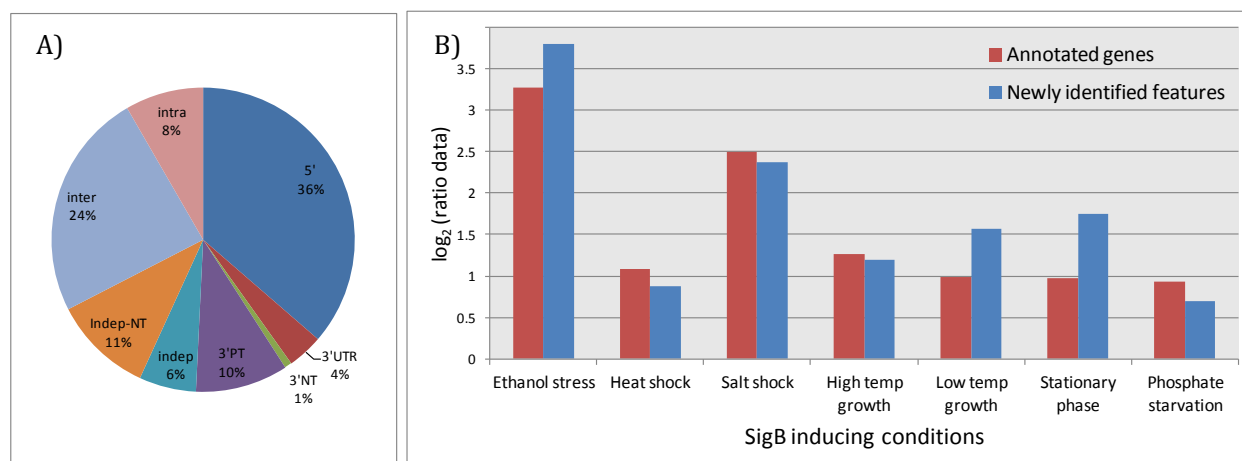
Fig. 13 A) Pie chart representing the distribution of the newly identified segments according to their structural relationships with annotated genes. (5' - 5' regions, 3'UTR- 3' untranslated regions, 3' NT- 3' extended transcripts with no termination, 3' PT - 3' extended transcripts with partial termination, indep - RNAs independent of previously annotated features with their own promoter and a termination site, indep-NT – RNAs independent of previously annotated features with no termination, inter - mRNA segment lying between two genes with distinct promoters, intra - portion of a polycistronic mRNA from a single promoter. B) Bar plot representing the average induction of annotated genes and the newly identified features. The induction profiles on the y-axis are expressed as the ratio of the expression value in the selected stress/growth condition to the expression value of the respective control, and are represented in log 2 scale. The various SigB-inducing stress / starvation conditions are shown in x-axis. The red colored bars represent the annotated genes, whereas the blue-colored bars represent the newly identified features in the tiling array study.

In general, the SigB activating stress conditions include environmental stress conditions like heat, ethanol, or salt stress, energy stresses like glucose or phosphate starvation, the entry into stationary phase and growth at high or low temperatures. In order to monitor the induction profiles of SigB-dependent genes, 40 SigB-dependent genes (Price, 2000) were selected and were termed as classical SigB genes in the entire thesis. Interestingly, this exploration of the data of the tiling array study revealed new conditions that induce the SigB regulon.
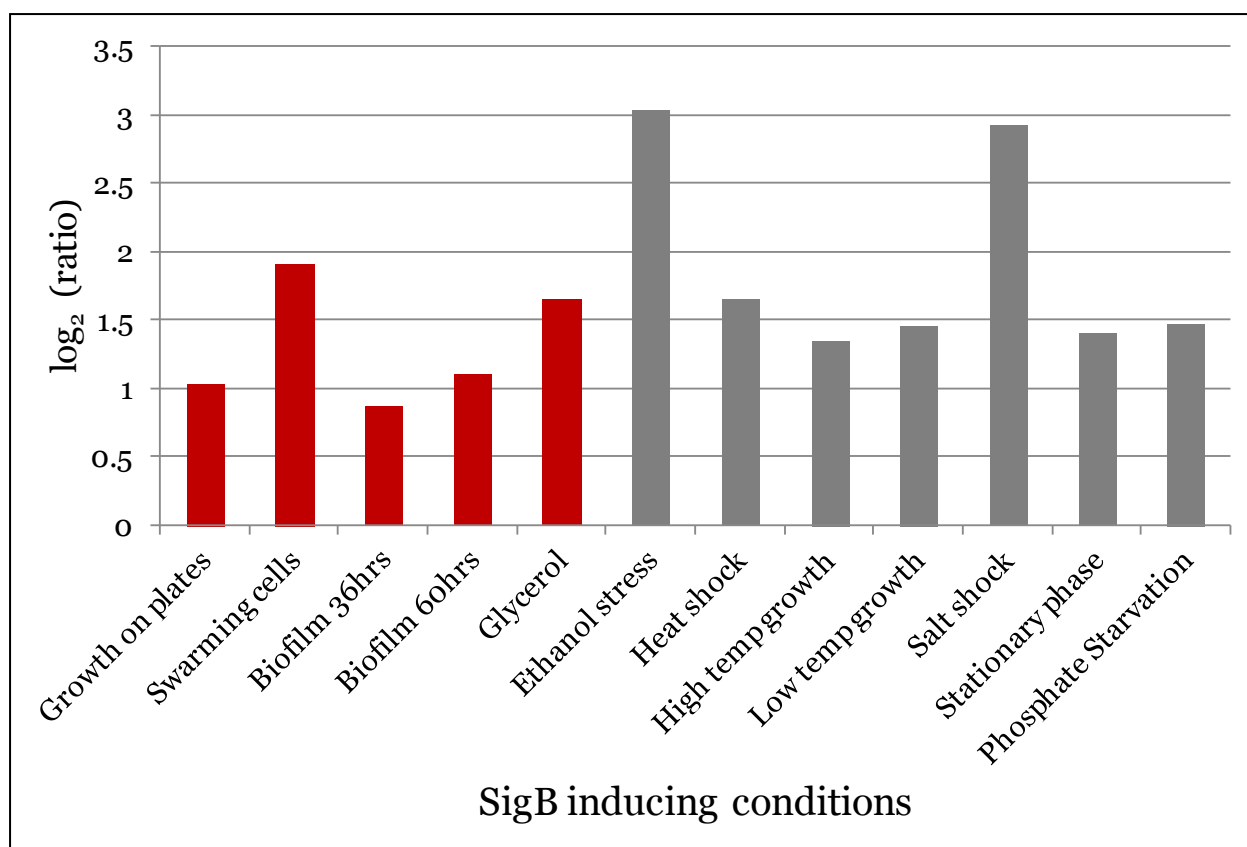
Fig. 14 Bar plot representing the average induction of classical SigB genes. The induction profiles on the y-axis are expressed as the ratio of the expression value in the selected stress/growth condition to the expression value of the respective control, and are represented in log 2 scale. The various SigB-inducing stress / starvation conditions are shown in x-axis. The grey colored bars represent the SigB-inducing conditions known so far, whereas the red-colored bars represent the conditions that are newly identified to be SigB-inducing conditions in the tiling array study.

Based on the expression of specific regulons under the newly identified conditions, the induction of SigB-dependent genes under the conditions of growth on plates and during swarming can most probably be attributed to nutrient limitation. During biofilm formation, the induction was increasing within the time interval from 36h to 60h, giving a hint that it might be due to the subsequent oxygen limitation. The underlying reason for the induction of the SigB regulon under glycerol has not been elucidated yet.

3.2.2 COMPARISON OF THE SIGB REGULON AS REVEALED BY THE TARGETED MICROARRAY STUDY AND THE TILING ARRAY STUDY

In the tiling array study, 415 segments including 133 new features were assigned to the SigB regulon (Nicolas *et al.*, 2012). As these targets were identified based on the expression data of the wild type, further validation using a *sigB* mutant might be useful.

As a first step we compared the list of 196 genes assigned to the SigB regulon based on the custom defined array by Random forest classification (Nannapaneni *et al.*, 2012) and the sigB regulon list obtained from the tilling array analysis. For the comparison, only annotated genes (282 genes) were taken into consideration. Comparison of the two studies revealed an overlap of 129 genes identified by both the studies. However, the selection of genes in the targeted microarray study was based on the previous transcriptome studies (Helmann *et al.*, 2001; Petersohn *et al.*, 2001; Price *et al.*, 2001) and thereby did not target the whole genome. Thus, the 153 genes identified only by the tiling array study highlight the potential of this genome-wide approach to identify new sigma factor targets (Fig. 15).
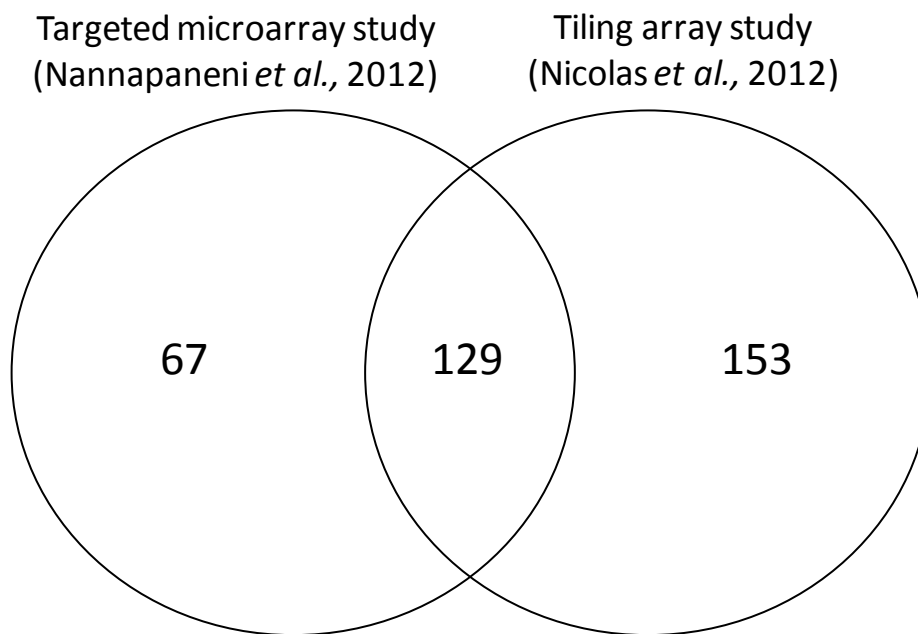
Fig. 15. Venn Diagram of the SigB regulon genes revealed by the two studies, the targeted microarray study and the tiling array study. The overlapping 129 genes represent the genes identified in both the studies, whereas the single lists represent genes identified by only one study. From the tiling array study, only annotated genes were considered for the comparison.

Among the 67 SigB-dependent genes only identified by the targeted microarray approach, one gene, *yjcE*, was no longer present in the new annotation (Barbe *et al.*, 2009). For seven of these genes, there is no SigB motif identified upstream. Among the remaining genes, 40 are the first genes of an operon and nine of these have good promoter scores (24%). The distribution of these genes in the four different clusters, which resulted from both expression and promoter ranking as described in the chapter 3.1.3, is shown in Fig. 16. The hierarchical clustering of the gene expression rankings to

the promoter scores in the targeted microarray study generated the clusters. Briefly, C1 corresponds to good promoter scores with relatively good expression rankings, whereas C2 corresponds to good promoter scores with slightly weak expression rankings.
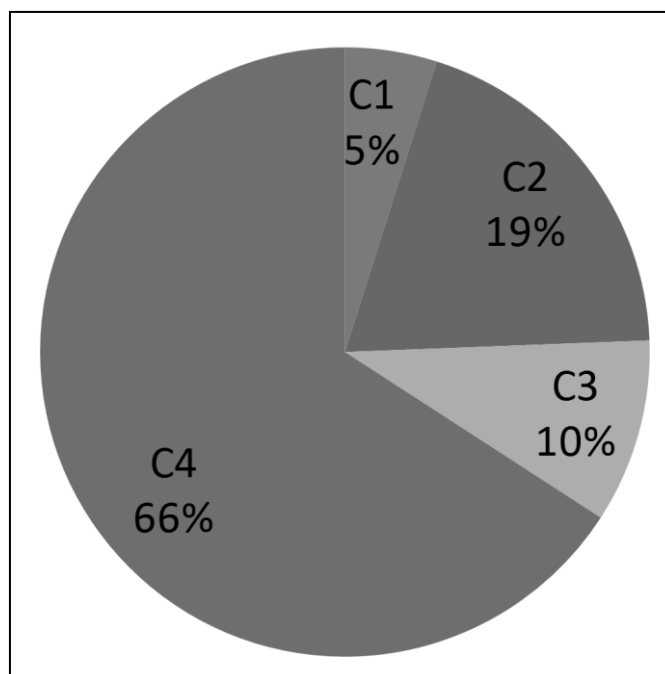
Fig. 16 Pie chart representing the distribution of SigB-dependent genes identified by only targeted microarray study. Among the 67 genes, the first genes of the operon (40 genes) were considered for this analysis. The percentage of genes in each cluster is shown in the figure. The clusters C1-C4 represent the clusters generated by the hierarchical clustering described in the chapter (3.1.3) of the thesis and shown in Fig. 10.

The average induction of these 40 genes in the SigB-inducing conditions from the tiling array is shown in Fig. 17. These genes were identified in the "expression RF" model (described in chapter 3.1.1 of the thesis), indicating significant differences among the wild type and its corresponding mutant strain. From the tiling array data, these genes still have significant induction under the conditions of ethanol stress, high temperature growth and salt shock (Fig. 17).
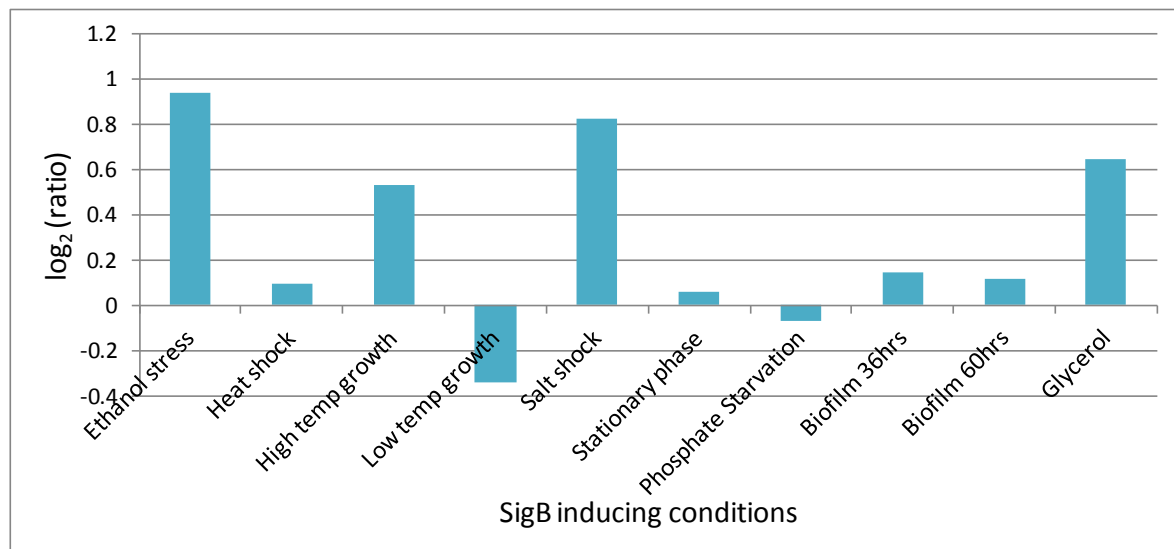
Fig. 17 Bar plot representing the average induction profile of the SigB-dependent genes identified exclusively by the targeted microarray approach. The induction profiles on the y-axis are expressed as the ratio of the expression value in the selected stress/growth condition to the expression value of the respective control, and are represented in log 2 scale. The various SigB-inducing stress / starvation conditions are shown in x-axis.

As shown in Fig. 15, 153 SigB-dependent genes were only identified by the tiling array study. These 153 SigB-dependent genes belong to 88 TUs. TUs were considered as known SigB-dependent TUs when either one or more genes in the TU belong to the SigB regulon from the targeted microarray study. Accordingly, 53 TUs comprising 90 genes were identified as new SigB-dependent TUs in the tiling array study. In order to examine the SigB-dependent expression pattern of these 90 genes, they were analyzed together with the set of 40 classical SigB genes (Price, 2000). These 130 genes (90 newly identified plus 40 classical SigB genes) were monitored under a set of known and newly identified SigB-inducing conditions and during sporulation. The genes were clustered based on the induction profiles into three different clusters by k-means clustering in the Rosetta Resolver software. The classical SigB genes were highlighted in the clusters (Fig. 18).
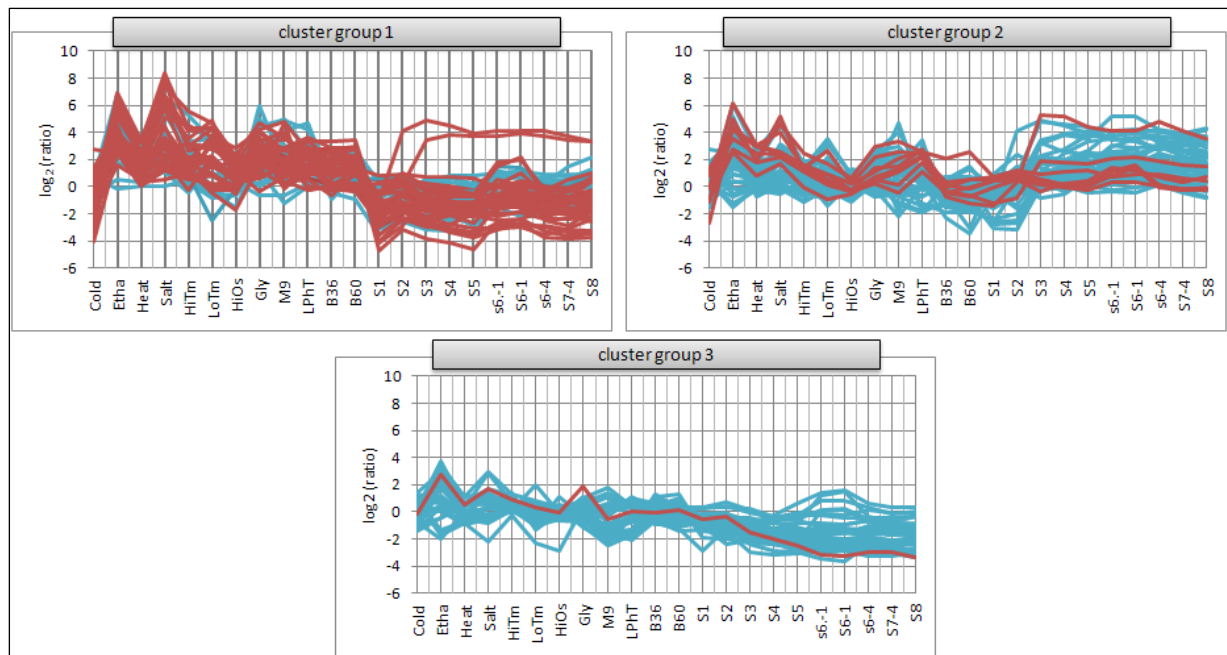
Fig. 18 K-means clustering of the induction values for the newly identified SigB-dependent genes (90 genes shown in blue color) from the tiling array study. Additionally, 40 classical SigB genes (Price, 2000) were included in the analysis (shown in red color). All the genes were assigned to the user-defined groups (three groups) using cosine correlation and sorted by its deviation. No additional data preparation or statistical cuts were applied. The ratio data set (experimental condition to its corresponding control condition) was represented in log 2 scale. Although not being continuous data, values were depicted as line graphs instead of bar charts to facilitate visual inspection. (cold – cold stress, etha – ethanol stress, heat – heat stress, salt – salt stress, high temp – high temperature growth, low temp – low temperature growth, high osmo- growth at high osmolarity, glycerol – glycerol stress, stat. – stationary phase, phosp starv - low phosphate starvation, biofilm 36, 60 - biofilm formation after 36h and 60h, spo 1, 2, 3, 4, 5, 6, 7, 8 – 1h, 2h, 3h, 4h, 5h, 6h, 7h, 8h respectively after induction of sporulation).

It is evident from Fig. 19 that the genes from clusters 1 and 2 mostly behave like classical SigB genes, i.e. have typical SigB-dependent induction patterns. When compared to the classical SigB genes, the 32 genes in cluster 3 exhibit different patterns, in particular lower induction under the SigB-activating conditions. The average induction levels of the cluster 3 genes along with the classical SigB genes are shown in Fig. 19.
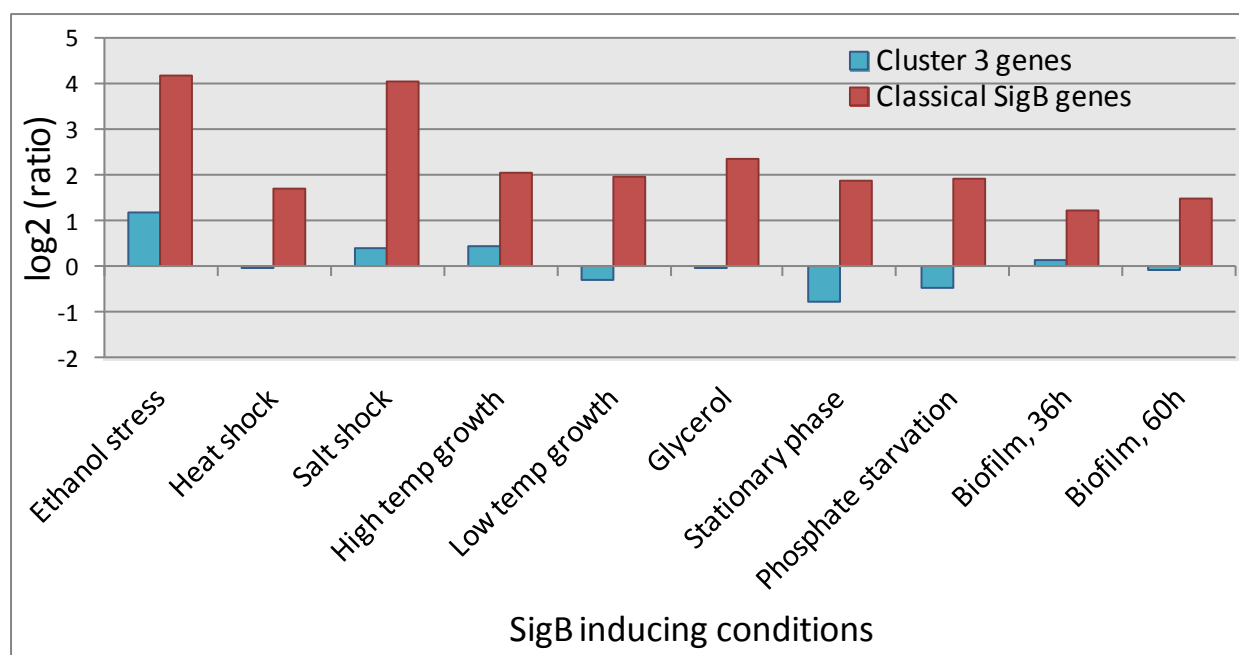
Fig. 19 The expression profiles of the cluster three genes (32 genes) under SigB-inducing conditions. Red-colored bars represent classical SigB genes showing a clear induction levels whereas the blue colored bars represents these cluster three genes that have no induction under the SigB-inducing conditions. The ratio data set (experimental condition to its corresponding control condition) was represented in log 2 scale.

These 32 genes have significantly lower induction levels, and in most of the conditions, they were not induced. Along with the SigB-dependent regulation, all these cluster 3 genes are regulated by at least one additional sigma factor, according to the tiling array study (Nicolas *et al* 2012, supplementary table S5). The effect of additional regulators controlling the expression of these genes might mask their SigB-dependent induction. The 32 cluster 3 genes with the corresponding regulators are listed in Table 2.

Table 2: The newly identified cluster 3 genes (32) with additional regulation

Gene[a]	BSU number	Description[b]	Sigma factor regulation[c]	Known secondary Regulator[b]
atpC	BSU36800	ATP synthase (subunit epsilon)	SigB, SigL, SigA	stringent response
cpgA	BSU15780	GTPase, activity stimulated by ribosomes, may be involved in ribosome maturation	SigB, SigA, SigA	
dgkA	BSU25310	undecaprenol kinase	SigA, SigB, SigGF	
folD	BSU24310	methylenetetrahydrofolate dehydrogenase (NADP)	SigA, SigB, SigA	PurR (repressor)
galE	BSU38860	UDP glucose 4-epimerase	SigA, SigB	
galK	BSU38200	galactokinase	SigB, SigA	CcpA regulon
galT	BSU38190	galactose-1-phosphate uridylyltransferase	SigB, SigA	CcpA regulon
ispA	BSU24280	Similar to an E. coli protein called IspA which is a farnasyl pyrophosphate synthase	SigA, SigB, SigA	
lytD	BSU35780	glucosaminidase	SigD, SigB	
menC	BSU30780	O-succinylbenzoate-CoA synthase	SigB, SigA	
nagBA	BSU35020	glucosamine-6-phosphate deaminase	SigB, SigA	NagR regulon
nusB	BSU24320	probable transcription termination	SigA, SigB, SigA	stringent response, PurR(repressor)
opuD	BSU30070	glycine betaine transporter	SigB, SigA	
rpe	BSU15790	ribulose 5-phosphate 3-epimerase	SigB, SigA, SigA	
tcyB	BSU03600	cystine ABC transporter (permease), membrane protein	SigA, SigB	
tcyC	BSU03590	cystine ABC transporter, ATP-binding protein	SigA, SigB	
thiN	BSU15800	thiamine pyrophosphokinase	SigK, SigB, SigA, SigA	
ung	BSU37970	uracil-DNA glycosylase	SigB, SigA	
xseA	BSU24300	similar to exodeoxyribonuclease VII (large subunit)	SigA, SigB, SigA	
xseB	BSU24290	similar to exodeoxyribonuclease VII (small subunit)	SigA, SigB, SigA	
yhcY	BSU09320	two-component sensor kinase	SigA, SigA, SigB	LiaR (activator)
yhcZ	BSU09330	two-component response regulator	SigA, SigA, SigB	LiaR (activator)
yhgE	BSU10160	similar to phage infection protein	SigB, SigA	
yqfF	BSU25330	unknown	SigA, SigB, SigGF	
yqfG	BSU25320	unknown	SigA, SigB, SigGF	
yqgC	BSU25030	unknown	SigB, SigA	
yqhY	BSU24330	unknown	SigB, SigA	
ytrP	BSU29650	two-component sensor kinase	SigA, SigB	
yutJ	BSU32200	putative NADH dehydrogenase	SigA, SigB	
yvoA	BSU35030	transcriptional regulator (GntR family)	SigB, SigA	
yvyI	BSU35790	mannose-6-phosphate isomerase	SigB, SigA	

Gene[a]	BSU number	Description[b]	Sigma factor regulation[c]	Known secondary Regulator[b]
yvzE	BSU35699	unknown	SigA, SigB	

[a] Genes are sorted based on gene name

[b] Descriptions and known secondary regulators are derived from subtiwiki
(http://subtiwiki.uni-goettingen.de)

[c] Sigma factor regulation was extracted from Nicolas *et al., 2012*. Mentioning SigA twice represents the gene is regulated by two SigA motifs upstream.

In the tiling array study, the predicted sigma factor binding site is reported for the detected up-shifts (supplementary table S4 of Nicolas *et al.*, 2012)*.* The SigB motifs were extracted for all the newly identified genes (90 genes) along with the classical SigB genes (40 genes) from the table. The obtained motifs were scored based on (Staden, 1984), for details see materials and methods (Appendix 3.2.2). Based on the scores for all these genes, a histogram was plotted in order to observe the frequency of the motif scores (Fig. 20). For the majority of newly identified genes, the frequencies of the motif scores are comparable with the motif scores of classical SigB genes. However, a part of the newly identified genes possess significantly lower scores (weak promoters).
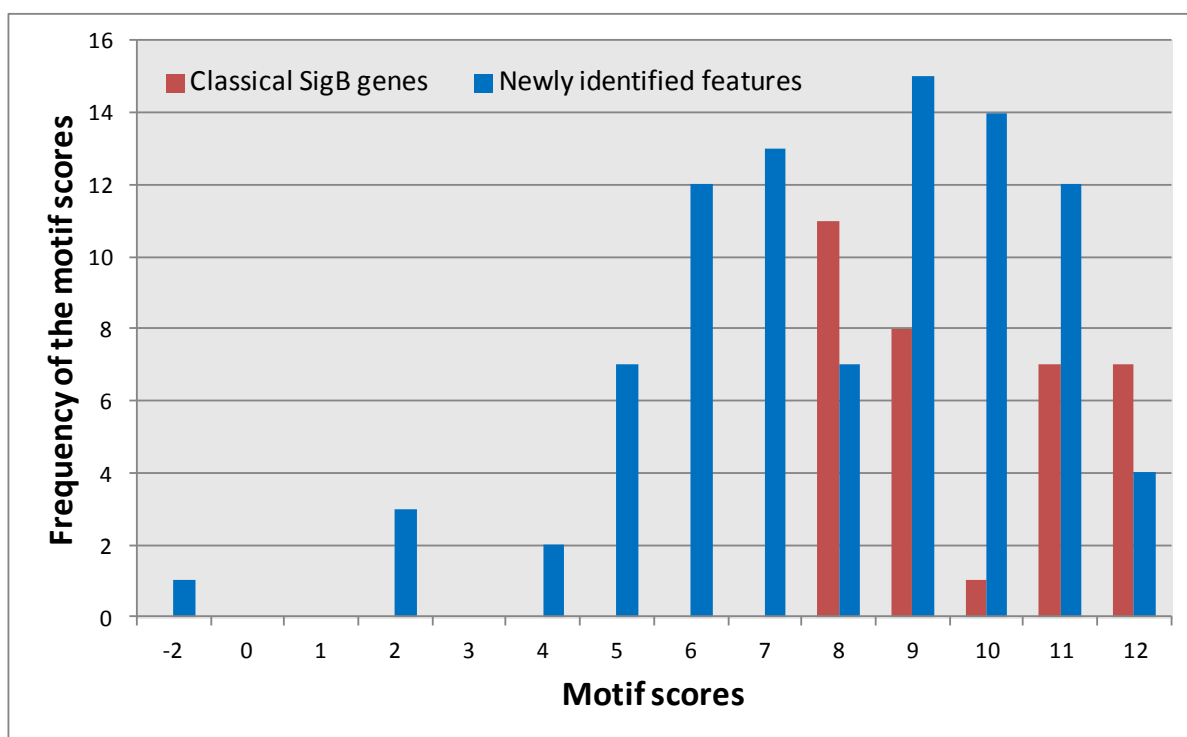
Fig. 20 Histogram representing the frequencies of the motif scores for the classical SigB genes as well as the newly identified features. Red-colored bars represent classical SigB genes whereas the blue colored bars represent the newly identified SigB features in the tiling array study. The x-axis represents the motif scores whereas the frequency of the motif scores was represented in y-axis.

3.2.3 NEW INSIGHTS INTO THE OSMOTIC REGULATION OF *BACILLUS SUBTILIS*

Among the wide variety of stress conditions, osmotic stress is one of the most common stress condition in the natural habitat of *B. subtilis*, as an increase and decrease of the water content of the soil is quite common. *B. subtilis* has an osmo specific adaptive mechanism during the conditions of salt shock and high osmolar growth. Additionally, the SigB-dependent general stress response complements this response. Among the proteins encoded by the osmo-adaptive genes there are three ATP dependent transport systems OpuA, OpuB, OpuC and two non-ATP dependent transport systems OpuD and OpuE which transport compatible solutes or its precursors into the cell. Besides this, the cell is capable to synthesize proline as osmoprotectant via ProH and ProJ, two alternative enzymes of the proline biosynthesis encoded by the osmotically induced *proHJ* operon. The potent osmoprotectant glycine betaine is either transported or synthesized via the enzymes GbsA and GbsB from choline. The corresponding osmo-adaptive genes were in general controlled by the sigma factor SigA. The genes *opuD* and *opuE* were additionally regulated by the alternative sigma factor SigB, and have a stringent SigB promoter motif upstream of these genes. The *proHJ* operon was additionally regulated by the sporulation sigma factor SigE.

A genome-wide transcriptome study of Steil *et al.* revealed the transcription of around 100 genes to be up-regulated with increase in osmolarity whereas the equal number of genes was simultaneously down-regulated (Steil *et al.*, 2003). A more recent study of Hahne *et al.* identified the osmotic up-regulation of about 500 genes during salt shock conditions (Hahne *et al.*, 2010). Besides, the number of genes identified by both these studies, the transcriptional profiling of the whole genome during these conditions was seen. Fig. 21 represents the comparison of these two studies to the tiling array expression ratios under the conditions of salt shock and growth at high osmolarity.

Fig. 21. A1)
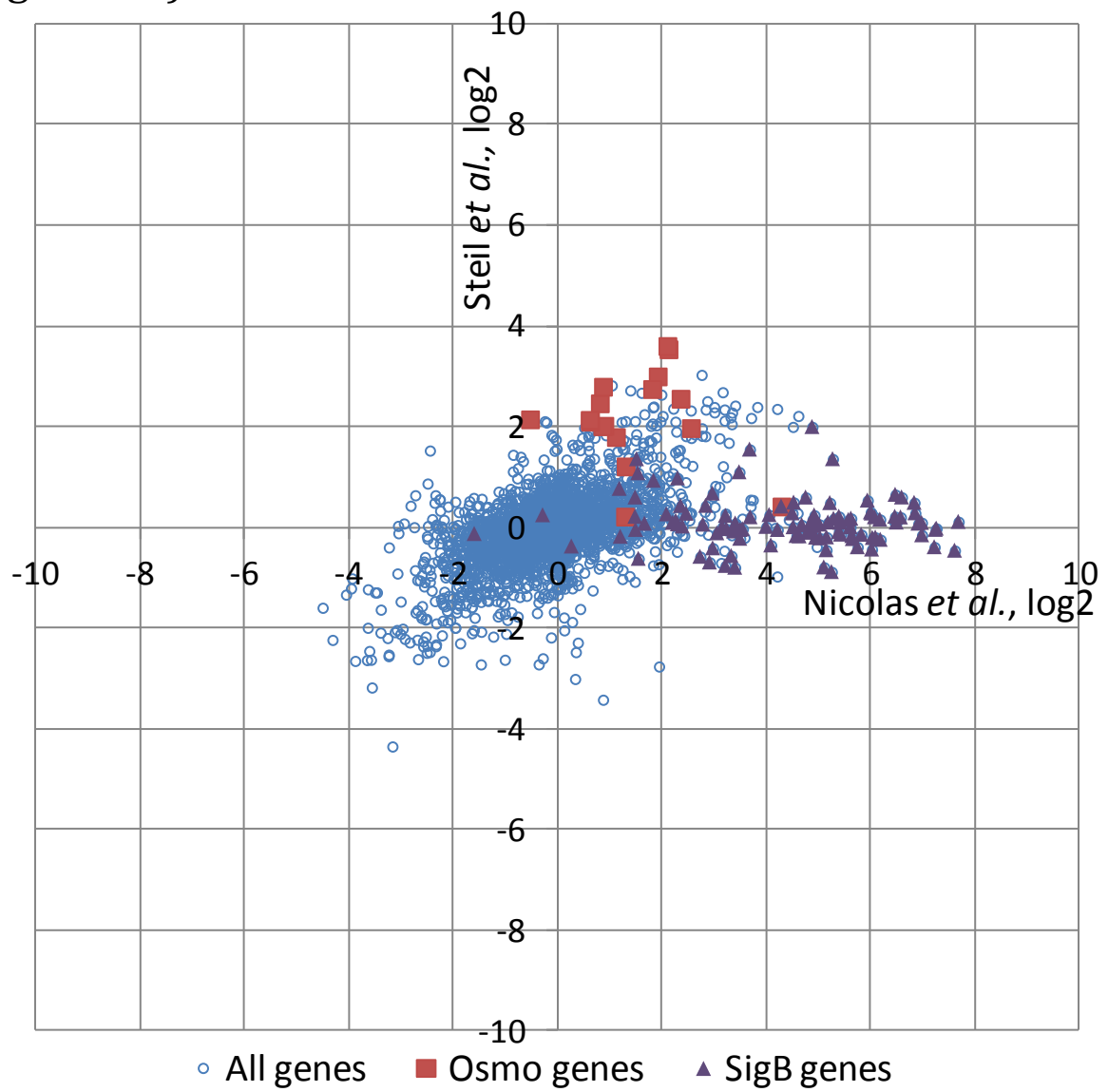
Fig. 21. A2)
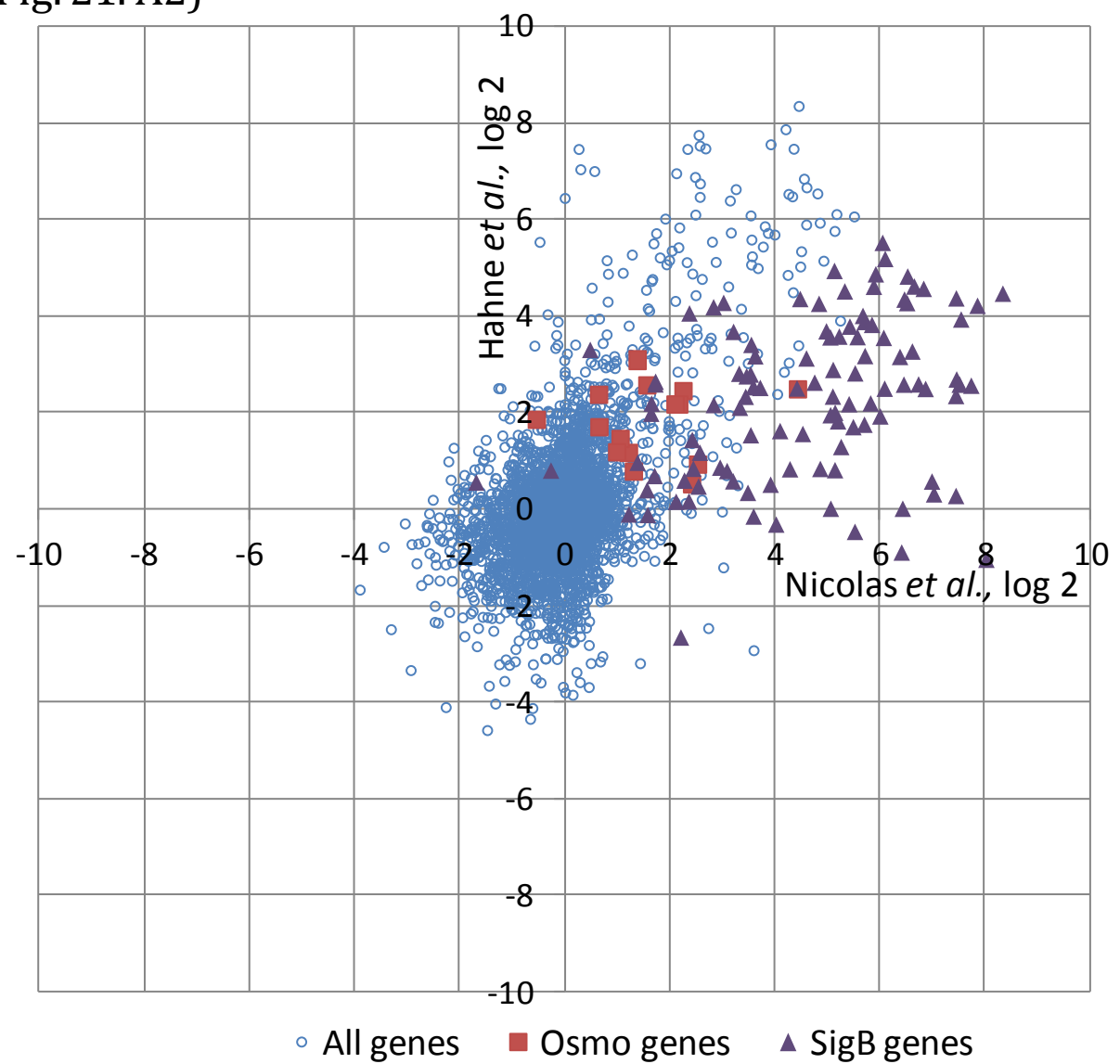
Fig. 21. A3)
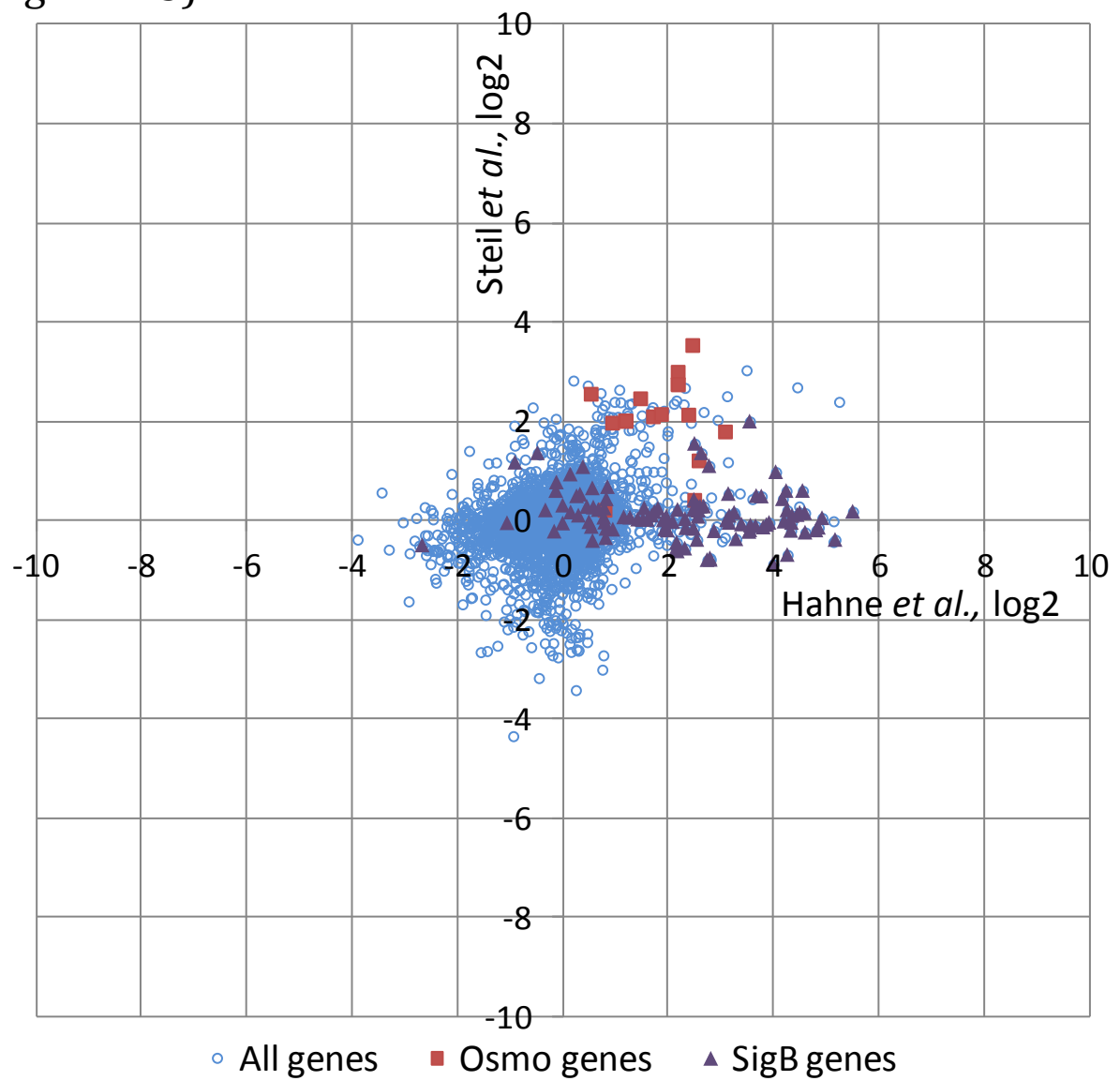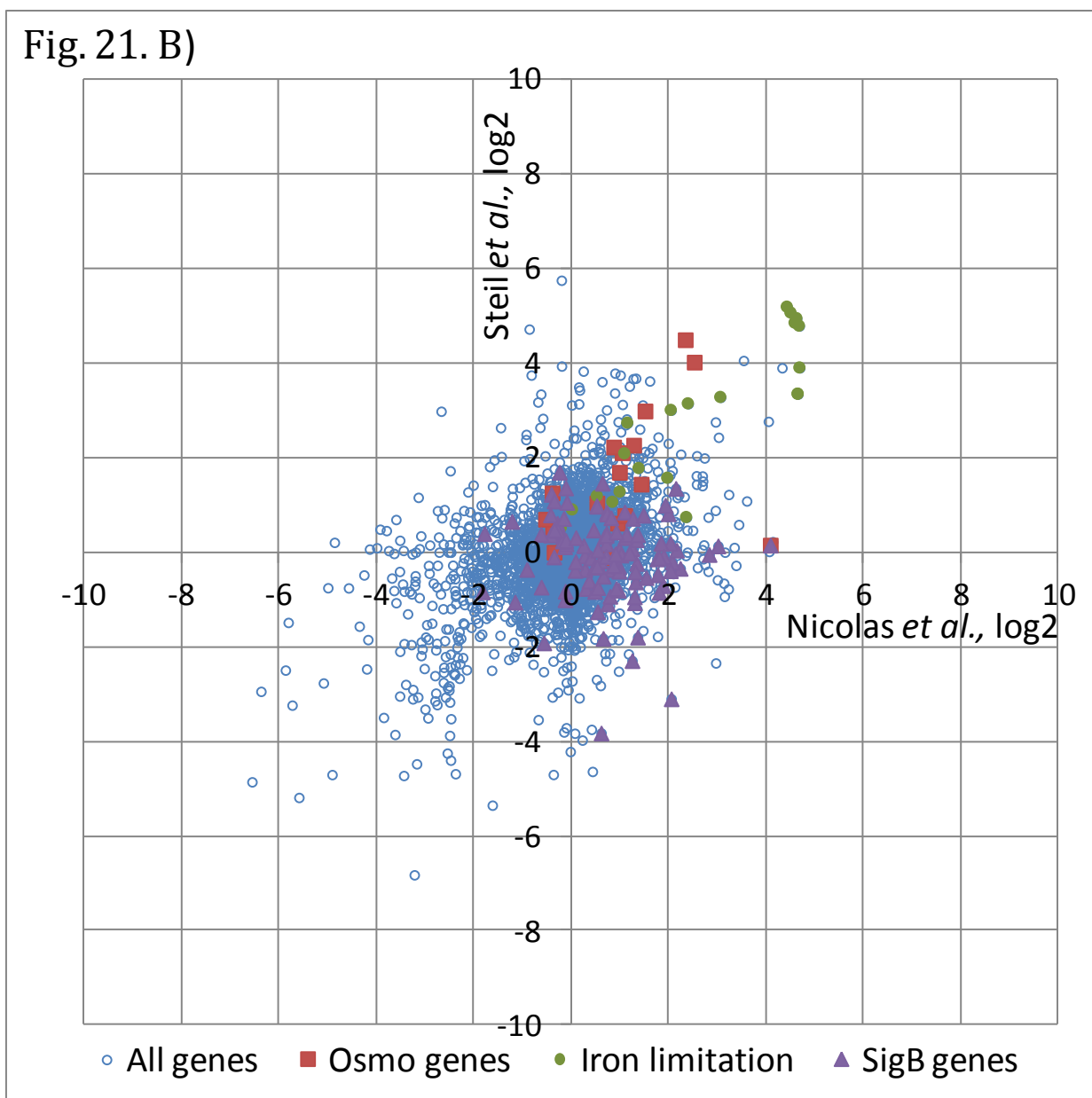
Fig. 21. B)

Fig. 21 A1-A3) The comparison of induction profiles of the genes from different data sets during salt shock. B) The comparison of induction profiles of the genes from different data sets during growth at high osmolarity. The log$_2$ values of the intensity ratios between the normalized signal intensities of the shock/high osmolar conditions and the signal intensities of control cultures are plotted. The intensity ratios of all genes were plotted in light blue dots in the background. The red squares represent the osmotically induced genes whereas the purple triangles represent SigB-dependent genes. The green circles represent the iron-limitation genes. The newly identified transcripts for the tiling array were not included in the analysis.

The intensity ratios of all genes were represented in the Fig. 21, whereas the osmo-adaptive genes and SigB-dependent genes were highlighted in red and purple colors respectively. The expression levels for all the genes were quite similar in all these studies, whereas the differences observed occurred mainly due to the different strains

used. Steil *et al* used the *sigB* mutant (BLOB22), a derivative of JH642 strain, Hahne *et al.* used the wild type 168 (t*rpC2*) and the tiling array analysis of Nicolas *et al.* used the BSB1 strain which is a tryptophan-prototrophic (Trp+) a derivative of the 168 *trpC2* strain.

The tiling array analysis allowed the identification of the maximal possible number of transcripts as it is independent of previous gene models. The median normalized gene expression intensities for the conditions of salt shock and high osmolar growth with its corresponding control condition were subjected to Welch's t-test and multiple testing correction (Benjamini and Hochberg). This enabled us to identify transcripts that are differentially regulated with $p < 0.05$, that also passed the fold change cutoff of two under osmotic response (Table. 3).

Table 3. Overview on the number of newly identified transcripts in the tiling array analysis.

	Salt shock		High osmolar growth	
	Induced	Repressed	Induced	Repressed
Newly identified transcripts	257	210	120	131

It is known, that the role of the Opu transporters extends well beyond the osmo protection of the cell, as they also mediate the uptake of the compatible solutes during low and high temperatures as well (Hoffmann & Bremer, 2011). Therefore, the tiling array data were used to get further insights into the expression of the osmo-adaptive genes under all the different growth and stress conditions tested in the study.

The tiling array technology provides data to identify new features controlling osmo-adaptive genes (antisense RNAs) and allows the elucidation of differences between the expression patterns of the *opu* genes. The next two chapters will highlight these two aspects of the tiling array analysis.

3.2.3.1 IDENTIFICATION OF NEW FEATURES CONTROLLING OSMO-ADAPTIVE GENES (ANTISENSE RNA 'S)

The tiling array technology monitors the transcription profiles of both DNA strands and allows the mapping based on signal up- and down-shifts to transcription units even if these were never reported before. Thereby an antisense RNA (*S1290*) was identified on the DNA strand opposite of the *opuB* operon. The antisense RNA covered the entire *opuB* operon, starting between *opuBD* and *opuBC*, as shown in Fig. 22A. In addition, a SigB-dependent promoter (class M11-from the unsupervised algorithm, which models the bipartite degenerate motifs) regulates the antisense RNA *S1290*. Also the induction levels of the antisense RNA under SigB-dependent conditions (Fig. 22B) indicated a clear SigB-dependent expression pattern in all the known inducing conditions. Due to the SigB-dependent antisense RNA covering the *opuB* operon, the induction of the *opuB* operon under salt shock condition, which also induces the alternative sigma factor SigB, was affected (Fig. 22C). Under salt shock, the expression of the antisense RNA was induced and the expression of the sense RNA (*opuB* operon, especially *opuBC*) did not show a significant induction (< 2-fold). Whereas under high osmolar growth conditions the expression of the antisense RNA was not induced, but expression of the *opuB* operon.
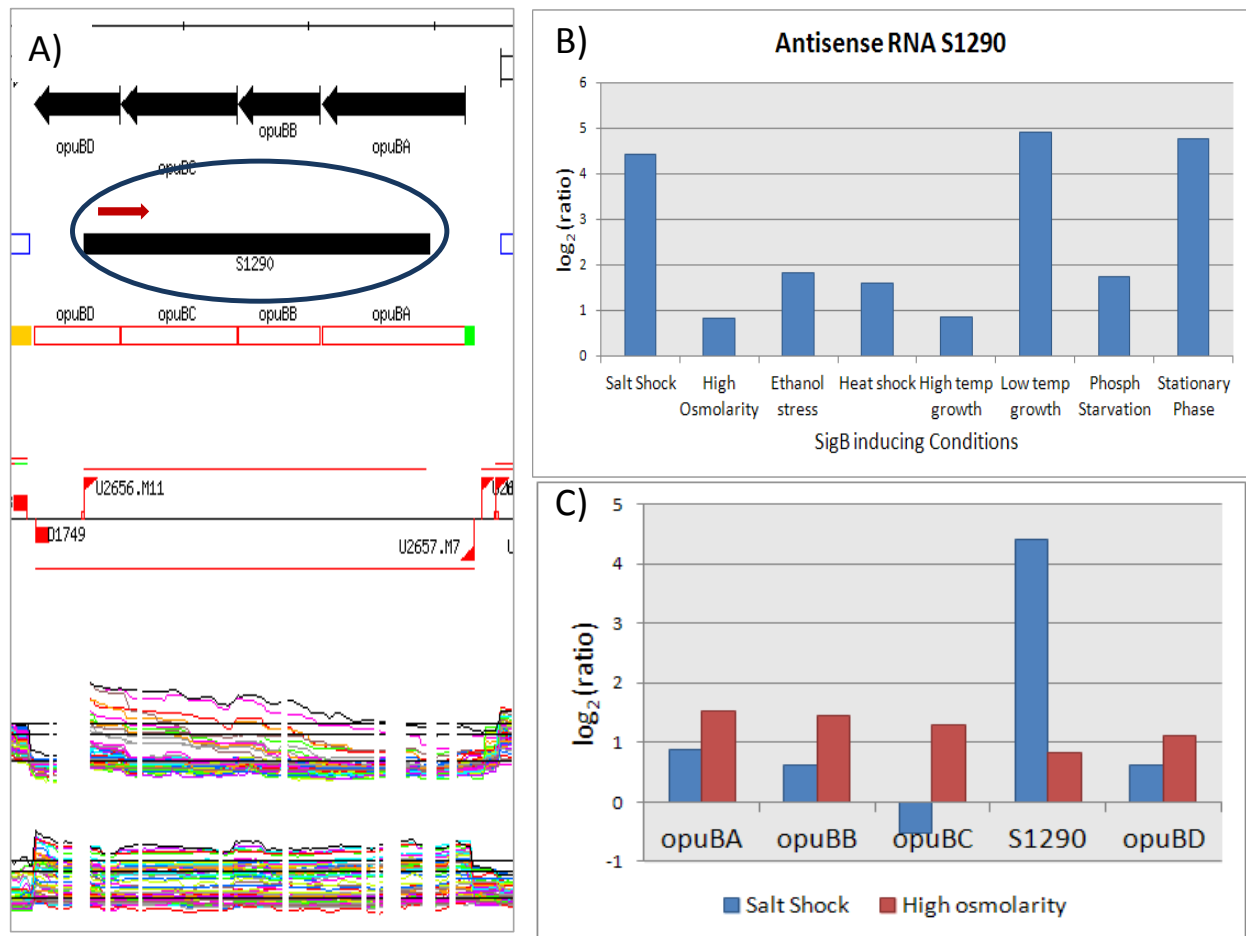
Fig. 22 A) Condition-dependent expression profiles of *opuB* operon extracted from Nicolas *et al.* 2012. The whole operon is under the control of SigA promoter (U2657.M7) upstream of *opuBA*. The antisense RNA *S1290*, present in the opposite strand of the *opuB* operon is under the control of a SigB promoter (U2656.M11). The start site of *S1290* is in between *opuBD* and *opuBC*. B) Bar plots representing the induction of the expression of the antisense RNA *S1290* under SigB-activating conditions. The induction pattern of the antisense RNA *S1290* clearly shows a SigB-dependent expression. C) Bar plots representing the expression profiles of the *opuB* operon under salt shock and high osmolar growth conditions. During salt shock the antisense RNA *S1290* showed a 20-fold induction compared to the control condition, thereby targeting the *opuBC* gene. During the continuous growth at high osmolarity, the *opuB* operon was not affected.

3.2.3.2 DIFFERENCES IN THE GENE EXPRESSION AMONG THE OSMO-ADAPTIVE GENES

The gene expression levels of the osmo-adaptive genes were monitored under a wide variety of conditions. In addition to the induction under osmotic conditions (salt shock and high osmolar growth) a significant induction or repression of the osmo-adaptive genes could be observed under conditions comparing growth in LB medium to growth in M9 minimal medium and at various time points during sporulation (Fig. 23). A detailed analysis under these conditions is shown in the later part of the thesis. In order to compare the level of induction of the two osmo-adaptive genes *opuD* and *opuE* that are additionally regulated by SigB, the osmo-adaptive genes were also examined under SigB inducing conditions. This analysis revealed that these two genes *opuD* and *opuE* exhibited different induction levels during most of the SigB inducing conditions. The lower repression levels of the osmo-adaptive genes was observed during the conditions of diamide stress, anaerobic growth conditions and growth in LB medium (Fig. 23).
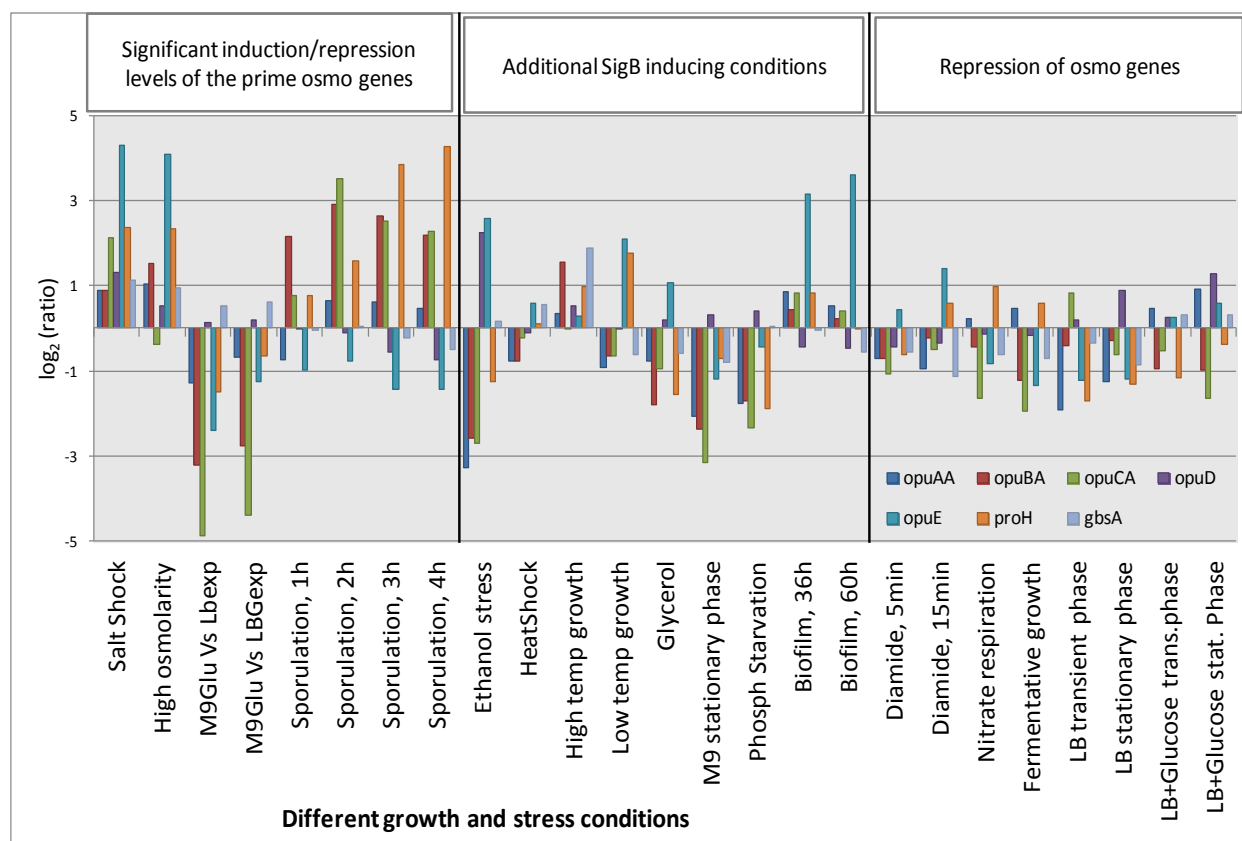
Fig. 23 The comprehensive picture of osmo-adaptive genes under different growth and stress conditions. The ratio data (experimental condition to its corresponding control condition) were represented in log 2 scale. The x-axis represents different growth and stress conditions As the genes in the operon behave similar, only the first genes of the operons were plotted in the figure. The picture was subdivided into three parts, where the first part represents the conditions where a significant induction or repression levels of osmo-adaptive genes was observed. The second part represents the induction levels of osmo-adaptive genes under SigB-inducing conditions and the third part represent the conditions where lower repression levels of osmo-adaptive genes were reported.

During the comparison of different growth media, we observed a strong repression of the osmo-adaptive genes in the LB medium compared to M9 minimal medium. This strong repression was also reproducible when the LB medium was supplemented with glucose (shown in Fig. 24). This repression could be observed for all osmo-adaptive genes, except for *opuD, gbsA, gbsB.*
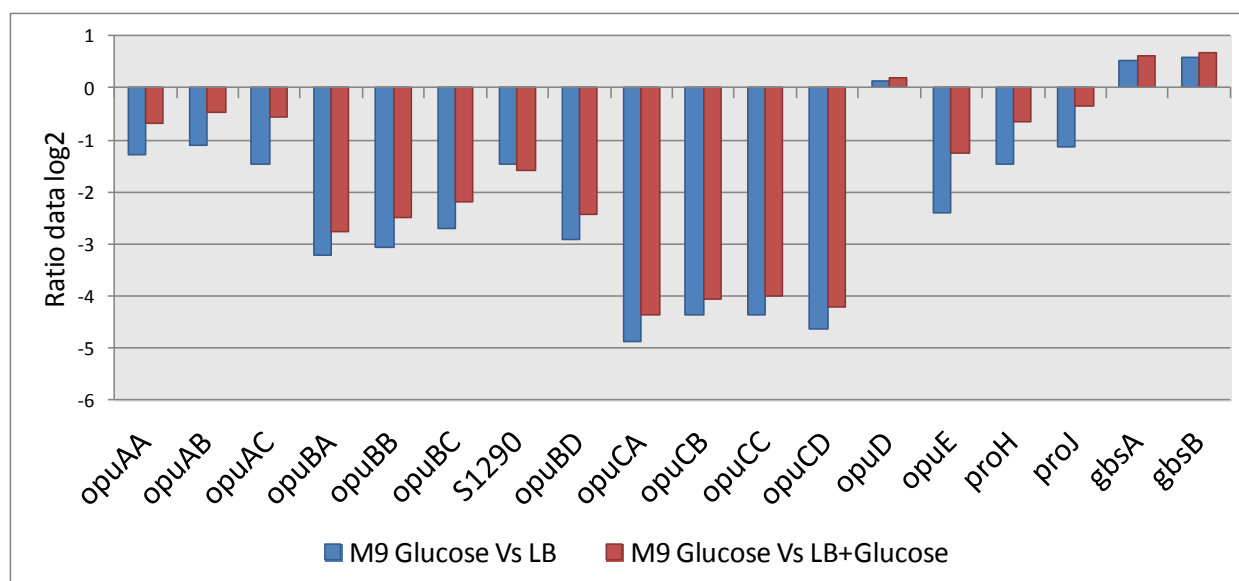
Fig. 24 Bar plots representing the repression of osmo-adaptive genes in LB medium with and without the addition of glucose. The ratio data set (experimental condition to its corresponding control condition) was represented in $\log_2$ scale. The blue colored bars represent the comparison of LB medium with respect to M9 medium, whereas the red colored bars represent the comparison of LB medium with glucose to M9 medium.

Another observation revealed by the inspection of the gene expression of the osmo-adaptive genes under non-osmo conditions was the induction of *proHJ* under sporulation conditions (Fig. 25A). This expression was likely driven from a SigE dependent promoter and started at two hours after the onset of sporulation (t2). During sporulation, the bacterial cell develops into a highly resistant spore, which is surrounded by a thick spore coat. As a high proportion of the proteins synthesized during sporulation are part of this spore coat, we inspected the percentage of proline within the spore coat proteins (Fig. 25B). The percentage of proline within the spore coat proteins was 5% and significantly higher compared to the percentage in all non-spore coat proteins 3.5%. This might indicate that proline and the alternative proline biosynthesis play a role during generation of resistant spores.
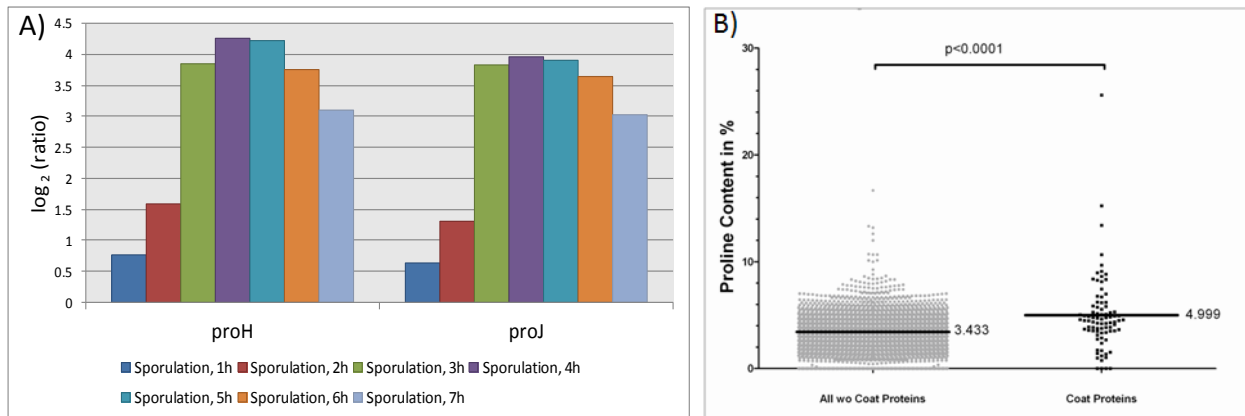
Fig. 25 A) Bar plots representing the induction of the *proHJ* operon during sporulation. Seven sporulation time points were shown 1h, 2h, 3h, 4h, 5h, 6h, 7h after the onset of sporulation. B) The plot showing the percentage of proline content in the proteins, divided into two sub-classes. The first class represents all the proteins excluding the spore coat proteins (labeled as All wo Coat Proteins) and the second class being the spore coat proteins (Coat Proteins). On the y-axis, proline content was represented in percentage.

In addition to the repression observed in LB growth medium (Fig. 24), a further repression of most of the osmo-adaptive genes was observed under diamide stress condition, where the cells were also grown in LB medium (shown in Fig. 26). The repression was reproducible in the two independent experiments, wherein the samples are taken at OD 0.6, 5min, 10min after the diamide stress and during OD 0.3 10min after the diamide stress. Almost all of the osmo-adaptive genes are repressed except for the SigB-dependent antisense RNA *S1290* and *opuE.*
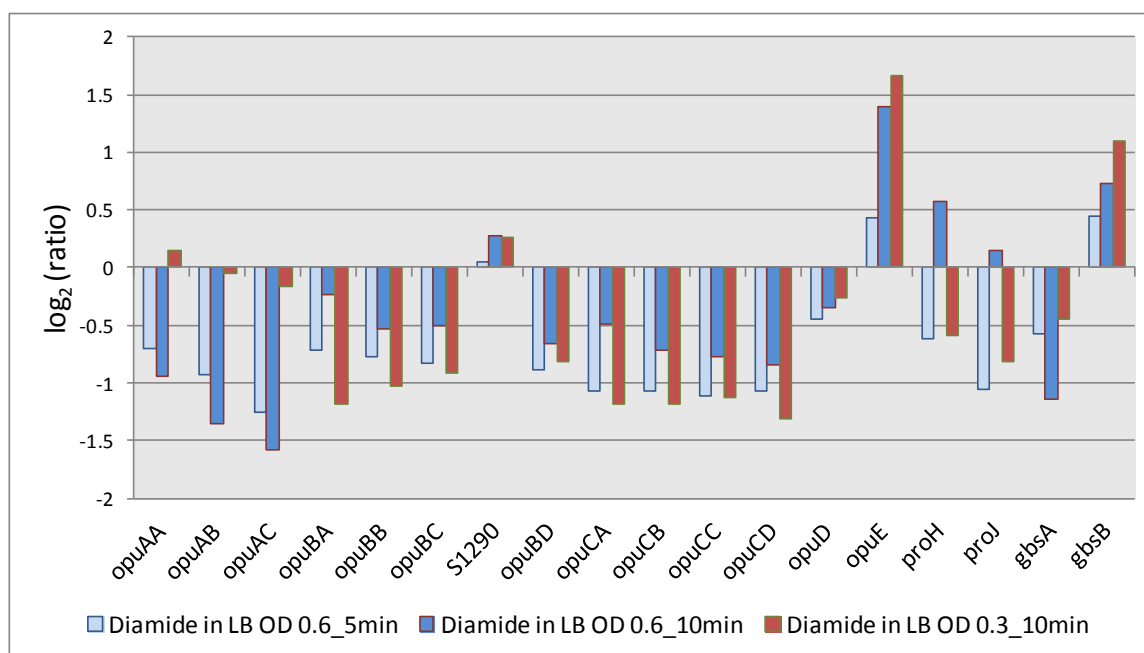
Fig. 26 Bar plots representing the repression of osmo-adaptive genes in diamide stress condition. The ratio data set (experimental condition to its corresponding control condition) was represented in $\log_2$ scale. The blue colored bars (light blue and dark blue) represent the diamide stress condition (5 min and 10min, OD 0.6) whereas the red colored bars represent the diamide stress condition (10 min, OD 0.3). The diamide stress was performed in the LB medium.

Because of the additional regulation of the anti sense RNA S1290, _opuD_ and _opuE_ by the alternative sigma factor SigB and the strong repression in LB media, we examined the expression of the osmo-adaptive genes under glucose starvation conditions. Therefore we used the data obtained from high-density fermentation experiments performed in M9 medium, where the cells were cultivated in the presence of glucose followed by glucose starvation. The time points before the exhaustion of glucose were represented as the negative time points and the time points after glucose was exhausted were represented as positive time points (Fig. 27). When the glucose was exhausted, the SigB-dependent general stress response was activated to cope with the starvation. As shown in Fig. 27, the repression of the osmo-adaptive genes starts immediately after the exhaustion of the glucose in the media. The differences of the expression pattern of the two SigB-dependent genes _opuD_ and _opuE_ was clearly visible as the _opuD_ expression remained stable throughout the starvation whereas the expression of _opuE_ was repressed. The expression of the SigB-dependent antisense RNA opposite of the _opuB_ operon strictly followed the transient SigB expression pattern (Fig. 27).
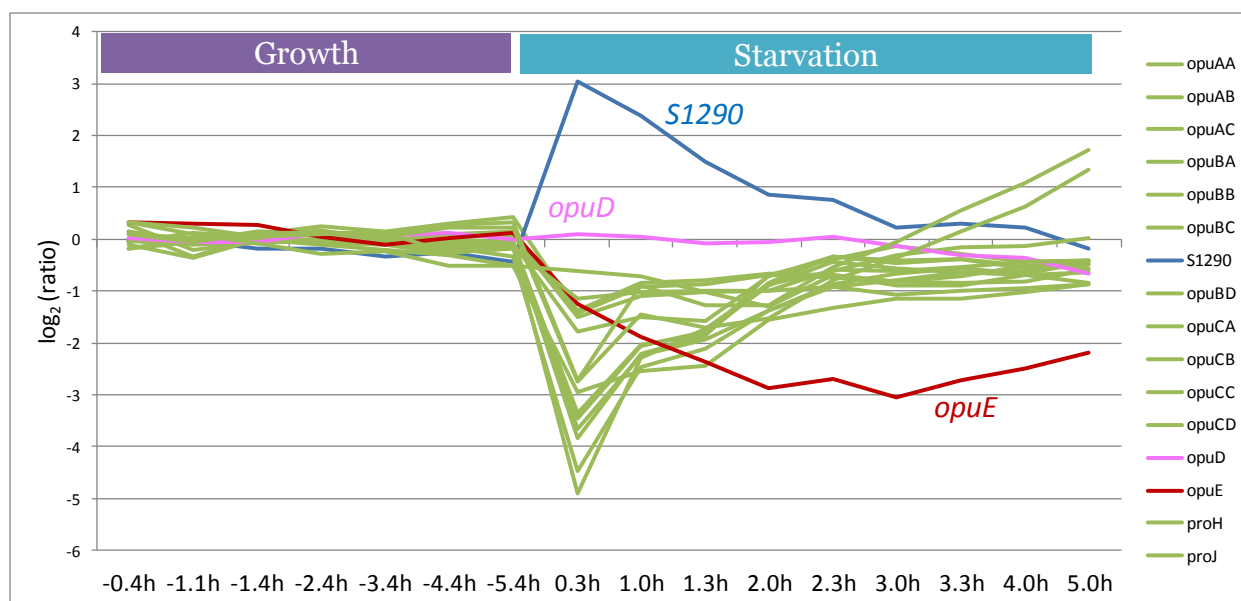
Fig. 27 The expression profiles of osmotically induced genes in the time series fermentation experiments, monitoring growth and starvation phases. As and when the cells enter stationary phase, the expression of the SigB-dependent antisense RNA *S1290* was drastically induced. The expression patterns of the osmo-adaptive genes *opuD* and *opuE,* which are additionally regulated by SigB, did not correlate with a classical SigB-dependent expression pattern. The blue colored line represents the antisense RNA S1290, targeting the *opuB* operon, the pink colored line represents *opuD* gene and the red colored line represents *opuE* gene. The cells are starved for glucose at the time point 0.0 h. The negative time points (-0.4 h, -1.1 h, -1.4 h, -2.4 h, -3.4 h, -4.4 h and -5.4 h) represents the sampling time points taken during growth phase whereas the positive time points (0.3 h, 1.0 h, 1.3 h, 2.0 h, 2.3 h, 3.0 h, 3.3 h, 4.0 h and 5.0 h) represents the sampling time points taken during starvation phase.

4. DISCUSSION AND CONCLUSIONS

The general stress response mediated by the alternative sigma factor SigB plays a key role in survival of *B. subtilis* during environmental stresses, energy limiting conditions and low temperature growth. Since the identification of SigB in 1980, the general stress response has been studied extensively (Akbar *et al.*, 1997; Antelmann *et al.*, 1996; Benson & Haldenwang, 1993; Brigulla *et al.*, 2003; Eymann *et al.*, 1996; Price, 2011; Völker *et al.*, 1994; Völker *et al.*, 1995). The genome wide identification of SigB regulon members revealed that the SigB regulon controls a large group of genes (Helmann *et al.*, 2001; Petersohn *et al.*, 2001; Price *et al.*, 2001). Although the general stress response has been studied for such a long time there are recent studies which give new insights into the function of the general stress proteins (Höper *et al.*, 2005; Reder *et al.*, 2012c) and define a sub-regulon within the general stress response mediated by MgsR (Reder *et al.*, 2008). In addition to these new aspects, the interconnection between the decision-making processes of the initiation of sporulation and the SigB-mediated stress adaptation were recently studied by Reder *et al.*, who provide new insights into the role of SigB during stationary-phase adaptation and long-time survival (Reder *et al.*, 2012a; Reder *et al.*, 2012b).

In order to understand the contribution of individual proteins to the stress adaptation, it is crucial to understand the structure of this large set of proteins induced during the general stress response. Therefore, the first step is the identification and characterization of the structure of the general stress regulon. When looking at the three major genome-wide transcriptome studies that were conducted to identify SigB regulon members, the resulting regulon lists have just 67 overlapping genes, although each of the study identified 100 to 200 regulon members (Helmann *et al.*, 2001; Petersohn *et al.*, 2001; Price *et al.*, 2001). Thus, the aim of this study was the definition of the structure of the SigB regulon using a targeted oligo nucleotide array, which was mainly derived from the regulon lists of the above mentioned studies, to compare the gene expression data of the wild type strain 168 and its isogenic *sigB* knockout mutant for almost all the SigB inducing conditions. In addition, we participated in an international consortium to analyze the expression levels of the wild type strain for a large number of conditions using high resolution tiling arrays, which added the layer of

mapped SigB-promoters and also resulted in genome-wide identification of additional SigB targets. Earlier studies defined the SigB regulon members using a small subset of defined stress conditions, the current work is the first to analyze the structure of SigB regulon with the majority of SigB inducing conditions. The identification of putative SigB-dependent promoter motifs, their correlation to gene expression in the targeted microarray study; the identification of SigB features within the newly annotated features and the identification of new conditions that induce parts of the SigB regulon in the tiling array study complement the work. The last part of the study focused on the osmotic response of *B. subtilis,* as new features could be identified that regulate the gene expression during the osmotic stress response.

4.1. DEFINING THE STRUCTURE OF THE GENERAL STRESS REGULON OF *BACILLUS SUBTILIS* USING TARGETED MICROARRAY ANALYSIS AND RANDOM FOREST CLASSIFICATION

In this study, *B. subtilis* wild type and its isogenic *sigB* mutant strains were subjected to a wide range of stress and starvation conditions. Rather than considering the induction by individual stresses, we exploited the broad activation spectrum of SigB. This turns into an advantage of this study over the previous transcriptional studies that addressed only few conditions while monitoring the general stress response. Thus, along with querying the global induction pattern of the SigB regulon under a variety of stress conditions, the single stress monitoring is always possible.

The identification of SigB-dependent genes by the three major genome-wide transcriptome studies revealed different sets of genes (Helmann *et al.*, 2001; Petersohn *et al.*, 2001; Price *et al.*, 2001) and the overlap among these gene sets remains quite low. This encouraged us to conduct the identification of SigB-dependent genes using a set of genes derived from these studies, rather than a whole-genome study, but use as much conditions known to induce the general stress response as possible. The introduction of suitable controls enabled the use of machine learning algorithms in this study, where a defined set of positives and negatives has to be considered.

The selection of the Random Forest algorithm among the different machine learning techniques is due to its ability to generate the feature importance and its wide range of applications for microarray data analysis (Diaz-Uriarte & Alvarez de Andres, 2006; Manilich *et al.*, 2011; Moorthy & Mohamad, 2011; Shi *et al.*, 2005; Wu *et al.*, 2008). The comparative analysis among the different machine learning algorithms revealed random forests having the highest accuracy. In terms of sensitivity and specificity, the random forest algorithm has also the better classification performance compared to support vector machines as well as the nearest neighbor algorithms (Yang, 2009). The sensitivity depends on the number of true-positives that are correctly classified, whereas the specificity depends on the number of true negatives that are correctly classified. The base classifiers that decide the classification in Random forests are the decision trees. The output of the decision trees are probabilities, and based on these probabilities the classification result can be justified. Decreasing this probability cut-off enables the trees to classify a record to positive, e.g. "SigB" class, to reduce the cost of false positives. However, this makes the random forests more precise but less sensitive.

Thus, instead of changing the probability cut-offs to decrease the number of false positives in the prediction, the training set is subjected to an ANOVA. This additional approach filters the training set based on the expression differences among the wild type and mutant strain, thus yielding a more stringent training set, and resulting in a stringent model. Thus, the implementation of ANOVA before training the models increases the sensitivity of random forests.

The combination of two random forest models, "expression RF" model and "kinetic RF" model enabled us to identify the SigB-dependent members with respect to the gene expression ratios as well as the identification of additional SigB-regulated genes possessing secondary regulators. However, the genes identified as targets for secondary regulators were validated by the presence of a SigB motif upstream of their transcriptional starting points. This validation has been necessary to prove the SigB-dependency of these genes.

4.1.1 IDENTIFICATION OF SIGB REGULON MEMBERS AND THE HEIRARCHY OF THE ANALYZED STRESS CONDITIONS

Earlier studies of Price *et al.* postulated 193 genes to be SigB regulon members by studying ethanol stress, 124 genes were assigned by the study of Helmann *et al.* using heat shock array data and Petersohn *et al.* identified a set of 124 genes to the SigB regulon, using ethanol stress, heat shock and osmotic stress. As these studies have a limited overlap of 67 genes, it was evident that the set of SigB-dependent genes varies based on the stress conditions studied. Additionally, the culture conditions and the strains used for the analysis might have contributed to the differences between the regulon gene lists established by these three studies. In our targeted microarray study, the SigB regulon members were assigned to the regulon by studying the majority of SigB inducing conditions – ethanol, heat, salt and oxidative stress, glucose and oxygen limitation, low temperature growth. The "expression RF" model identified 166 regulon members based on the wild type to mutant expression ratios. The list of SigB regulon members was further extended by analyzing the additional regulators by the "kinetic RF" model.

Along with the identification of SigB regulon members, the RF has allowed to rank the stress conditions used in the study based on their influence in the classification. The ranking of the stress conditions indicates that the RsbU-dependent environmental stresses play a very important role, and that these conditions are much more potent to trigger the general stress response compared to the RsbP-dependent energy stresses (Fig. 8). Within the stress conditions, the importance of the time points to classify SigB dependence follows the transient nature of SigB-dependent genes (Völker *et al.*, 1995). Oxygen limitation and oxidative stress play a minor role in the activation of the general stress response. Even though a recent study of Reder and co-workers revealed that a substantial number of general stress genes is involved in the development of oxidative stress resistance (Reder *et al.*, 2012c).

The second RF model, "kinetic RF", predicted a set of 37 genes potentially controlled by an additional regulator, 19 of these genes have an additional regulator according to SubtiWiki. In addition, the genes were searched for additional regulators in the tiling array analysis as a further confirmation of the literature data, indicating 34 of these 37

genes having a regulator additional to SigB. This supports that the expression of these genes in the mutant strain most probably more or less completely complements the SigB deficiency. As these genes were predicted solely by the expression data in the wild type strain, an additional validation is necessary to prove the SigB-dependency of these genes. The presence of a putative SigB promoter motif upstream of these genes proves the SigB-dependency. Among the 37 genes, seven genes were not preceded by a putative SigB motif, and therefore these genes were not included in the final set of the SigB-dependent genes.

Furthermore, 15 genes out of the 166 genes predicted by the "expression RF" model were not identified by the "kinetic RF". Due to their low expression levels in the wild type strain, these genes were not identified by the "kinetic RF" model, which classifies only based on the wild type expression data. As the expression levels of these 15 genes were lower in the mutant strain, the resulting induction ratios were still high enough leading to the prediction of these genes to the "SigB" class by the "expression RF" model. Therefore, these 15 genes were included in the final list of SigB-dependent genes. Thus, 196 genes could be assigned to the list of SigB-dependent genes that are induced in all the stress conditions tested in the study.

The motif search was performed for the region upstream of the identified SigB-dependent genes, the obtained motifs were scored, and then these were combined with the expression rankings, and simultaneously subjected to hierarchical clustering. The resultant clusters were analyzed for the mediated regulatory changes. As expected, the cluster C1 contains the majority of the experimentally verified promoters. Cluster C2 contains genes having good promoter scores and lower expression ratios. Six of the 20 genes within this cluster C2 were subjected to additional regulation according to SubtiWiki (Florez *et al.*, 2009). Cluster C3 consists of genes having low promoter scores and high expression ratios. This cluster C3 together with C4 contains fewer genes having an experimentally verified promoter, compared to the remaining clusters. Based on this analysis we conclude that there are genes having a good correlation between expression ranks and promoter scores in cluster C1 and C4, which cannot be generalized as the genes in cluster C2 and C3 do not show this strong correlation of promoter stringency and expression values.

The SigB regulon comprising 196 genes identified in the study is also supported by the results of an extensive phenotypic screening performed by Höper *et al.* (2005). This study identified stress sensitivity of 94 mutants of candidate SigB regulon members using a 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 analyzed, 80 were sensitive to at least one of the four stress conditions tested. In the targeted microarray 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* showed the strongest differences to the wild type strain during low temperature growth. In our analysis, these genes were induced under all the conditions tested in the study, along with low temperature growth. This example also highlights the value of the study, because these genes were previously identified only by Petersohn *et al.* (2001) and would have been missed if the focus would be only on the 67 overlapping genes.

Supporting previous studies, the targeted microarray study revealed that oxidative stress contributes least to the induction of the general stress response (see hierarchy plot of the variable importance in Fig. 8). A recent study from Reder and co-workers (Reder *et al.*, 2012) analyzed the contribution of the SigB regulon to the secondary oxidative stress resistance of *B. subtilis*. As the oxidative stress is the most common secondary component for many of the primary SigB inducing stimuli like ethanol, salt or cold stress or a combination of these stresses (Höper *et al.*, 2005), the secondary oxidative stress was investigated. Therefore, wild type strain and the cells of the 94 mutants, which were previously analyzed by Höper, were pre-adapted for 20 minutes using ethanol, allowing the cells to synthesize general stress proteins. Thereafter oxidative stress (paraquat or H_2O_2) was applied to the cells of the different mutants. The analysis validated the key role of the SigB-dependent general stress response and identified individual mutants that are more sensitive to secondary oxidative stress compared to the pre-adapted wild type strain.

Around 95% of genes, which were described to be 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 our study (Fig. 11). The studies of Price *et al.* and Helmann *et al.* were single stress specific, and the genes identified by only one of these studies clearly indicate stress specific genes rather than the global monitoring of the general stress response. Thus, for genes reported to be SigB-dependent in only one of these studies the frequency of validation by the targeted microarray study was clearly lower. Of the 96 candidate genes only identified by Price *et al.* (2001), 62 genes were not identified as SigB in our study (Fig.11). Surprisingly 60% of these genes have a predicted SigB promoter, despite the fact that they did not show an induction in most of the tested stress conditions except ethanol stress. Thus, the specific induction of most of these genes (90% of these 62 genes) in the targeted micro array study only after ethanol stress likely reflects a thus far unidentified ethanol specific regulation deviating from the typical SigB-dependent induction by other stimuli. Thus, these 62 genes can be assigned to ethanol-specific regulated genes. Similarly, a small sub-set of 11 genes identified by Helmann *et al.* (2001) was not identified in our study, though they exhibit induction levels under heat shock conditions.

In conclusion, the current work on the targeted microarray study presents a systematic identification of SigB regulon members using two random forest models named "expression RF" and "kinetic RF", based on a comprehensive microarray where most of the SigB inducing stress conditions were queried. The targeted analysis used in this study covered all potential SigB-dependent candidate genes according to previous knowledge. Inclusion of environmental and energy stresses as well as low temperature growth also ensured that prospective SigB targets indeed displayed a broad induction pattern, which should be an important determinant for assigning a gene to the regulon. The stress hierarchy predicted in this study revealed environmental stresses having the ability to trigger the general stress response more effectively compared to the energy limiting stress conditions. Integrating the observations of previous studies with those of the current study, we defined the SigB regulon with 196 members. This knowledge can serve in further studies to reveal the specific functions of these genes in *B. subtilis* stress management.

4.2 *BACILLUS SUBTILIS* WHOLE TRANSCRIPTOME ANALYSIS USING TILING ARRAYS

The tiling array analysis covered the majority of conditions that the bacterium can encounter allowing to identify and map the maximal possible number of transcription units to its corresponding sigma factor regulons. Along with the identification of several newly identified transcription units, the analysis unraveled that sigma factor utilization accounting for ~66% of the variance in transcriptional activity, is mainly associated with survival and adaptation processes like sporulation and stress responses. Additionally the highest proportion of variation attributed to sigma factor activity was obtained for the general stress sigma factor SigB, the chemotaxis sigma factor SigD and the four sporulation sigma factors. This indicates that condition specific regulation of gene expression is substantially based on specific sigma factors. In the current study the tiling array data were utilized to extend the targeted microarray study of SigB regulon members to a whole genome-scale and to examine the regulatory mechanisms of the osmotic stress response with respect to new features and conditions.

With the inclusion of several stress and starvation conditions (ethanol stress, heat shock, salt shock, high and low temperature growth, high osmolar growth, stationary phase growth and phosphate starvation) on the array and the presence of grouped SigB promoters identified in the study, 415 genes and RNA features were assigned to the SigB regulon. After the subtraction of the 282 already annotated genes of *B. subtilis,* 133 newly identified RNA features including a large number of antisense RNAs were observed. Further studies using techniques like high throughput transcriptome sequencing are required to analyze this group of features.

4.2.1 IDENTIFICATION OF SIGB-REGULON MEMBERS

To investigate the varying induction levels in the tiling array data within the SigB regulon, the annotated 282 SigB-dependent genes from the tiling array study were compared to the oligonucleotide spotted microarrays. The comparison revealed an overlap of 129 genes, which were assigned to the SigB regulon by both studies. The remaining 153 genes were only identified in the tiling arrays because the majority (138 genes, 90%) of them was missing on the targeted arrays. Out of these 153 genes, a set of 90 genes could be created, comprising transcription units not known to be SigB-

dependent before. Further analysis of this set of newly identified SigB-dependent genes together with classical SigB genes (40) by k-means clustering based on their induction patterns revealed a cluster of genes (cluster group 3, Fig. 18) that has low induction levels in all the tested stress conditions. Even though a part of the genes possesses significantly lower motif scores, the majority of genes in this cluster possess stringent SigB promoter motifs upstream (Fig. 20). The presence of at least one additional regulator for this group of genes might dampen the SigB–dependent induction levels. However, the presence of a stringent SigB motif for the majority of these genes (32 genes) of the cluster group 3 (Table 2), which was also validated by a Bioprospector search formulate these genes as considerable subjects of interest for further studies.

4.2.2 FUNCTIONAL ANALYSIS OF THE SIGB-DEPENDENT GENES

The k-means clustering of the whole SigB regulon was performed based on the gene expression data during SigB inducing conditions to check the enrichment of single clusters having specific regulators. However, the genes regulated by these sub-regulators like MgsR and NsrR, do not really differ compared to the expression levels of other SigB-dependent genes, and thereby none of these was identified in the targeted array data as well as the tiling array analysis. It is known from the previous studies and the current work that the function of the majority of SigB-dependent genes is still unknown (Hecker *et al.*, 2007; Price, 2002). The functional analysis of the newly identified SigB regulon members of the tiling array analysis (153 genes) revealed that the majority of those genes code for GTP-binding proteins or are related to RNA synthesis and degradation (Fig. 28). It is known from the previous studies that *obg*, encoding a *B. subtilis* GTP binding protein, is essential for growth, the activation of general stress regulon and sporulation (Kok *et al.*, 1994; Scott & Haldenwang, 1999; Trach & Hoch, 1989; Vidwans *et al.*, 1995). The over-representation of GTP-binding proteins among the newly identified regulon members links the role of such regulators like Obg to the SigB-dependent general stress response. Therefore, this functional classification can serve as input for further studies.
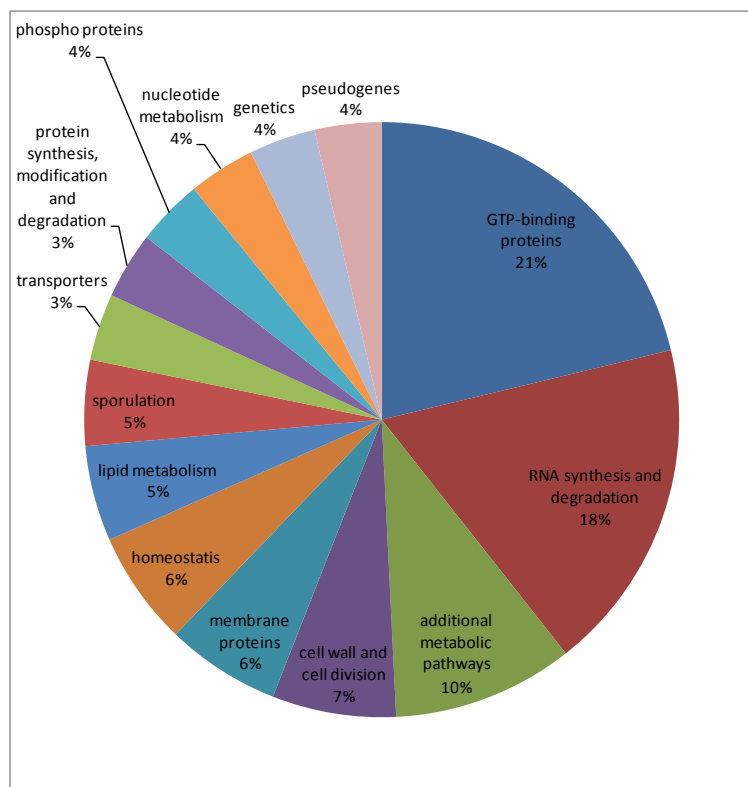
Fig. 28. Pie chart representing the distribution of the functions for the newly identified SigB-dependent genes (153 genes) in the tiling array study. The functions for the genes were extracted from SubtiWiki. The name of the function and their distribution was labeled on the figure.

4.2.3 NEW CONDITIONS THAT INDUCE SIGB REGULON

The analysis of the general stress response based on the tiling array data revealed the induction of SigB-dependent genes under a set of conditions like growth on plates, swarming cells, biofilm formation and growth on glycerol as carbon source which do not belong to the classical SigB inducing conditions. To analyze the underlying reasons for this activation of the SigB-dependent gene expression, the specific regulons expressed under these conditions were examined.

During growth on plates as confluent colonies and as swarming cells, the most relevant stress conditions are nutrient limitations due to limiting amounts of glucose, oxygen and phosphorus. This was evident from the up-regulation of these respective regulons from the tiling array data. The tiling array data demonstrate that cells face severe glucose limitation when grown on plates, as genes controlled by the CcpA repressor are up regulated (Fig. 29 D). CcpA, the global transcription regulator is

involved in glucose regulation of a large set of genes, by repressing the gene expression of catabolic genes in the absence of glucose (Fujita, 2009; Henkin, 1996; Singh *et al.*, 2008; Tobisch *et al.*, 1999). The most probable reason for the activation of SigB-dependent general stress response might be the entry of cells into stationary phase due to subsequent glucose limitation and is evident from the parallel up regulation of SigH regulon in both of the conditions, growth on plates and swarming cells (Fig. 29D, 29E).
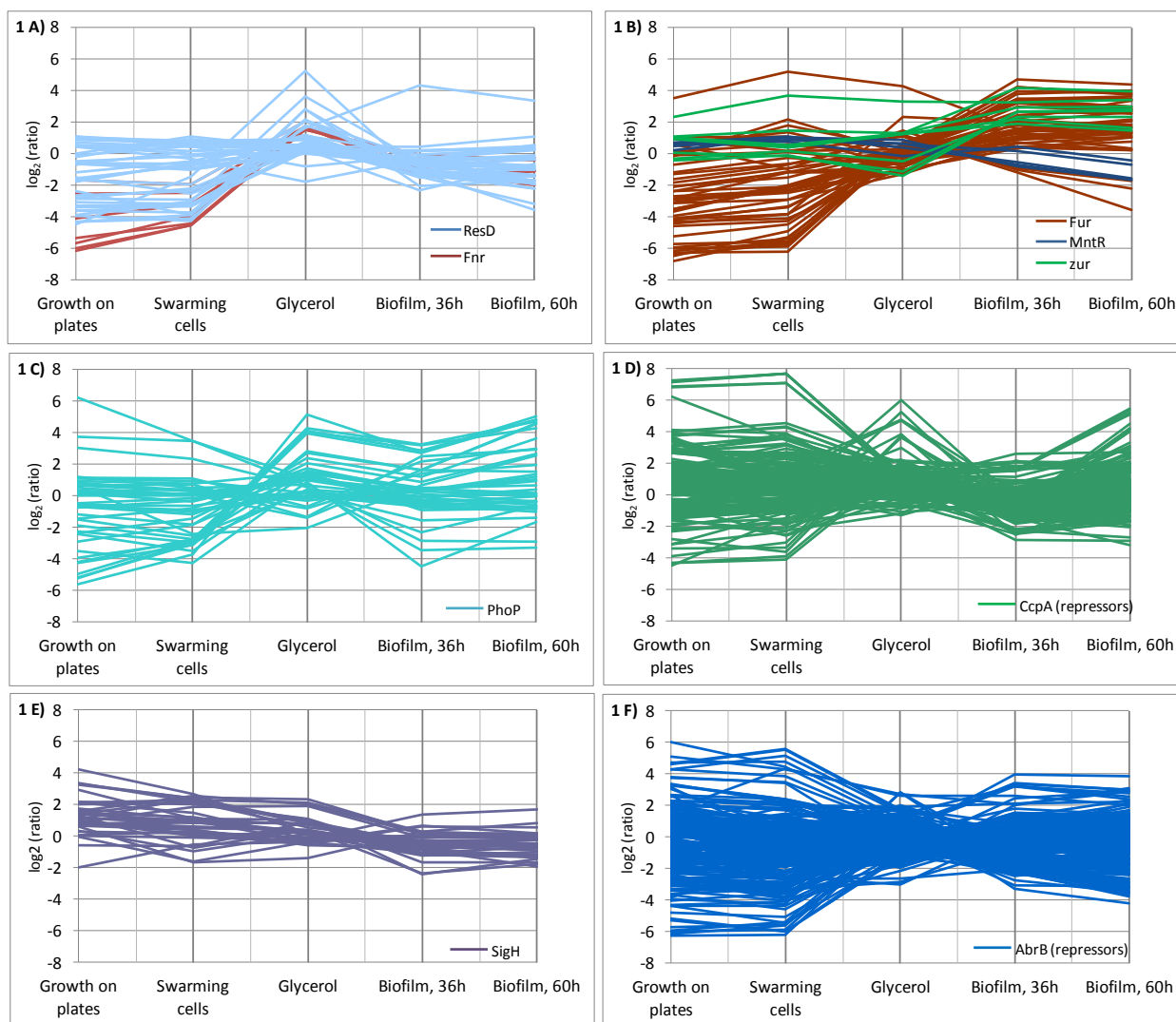
Fig. 29. The induction levels of the specific regulators / sigma factors (labeled on each plot) in the newly identified SigB inducing conditions. Although not being continuous data, values were depicted as line graphs instead of bar charts to facilitate visual inspection. The log$_2$ values of the intensity ratios between the normalized signal intensities of the respective stress conditions with the signal intensities of control cultures are plotted.

The experiments that are associated with the change of the carbon source to glycerol instead of glucose revealed a strong induction of all the SigB-dependent genes. While monitoring the different regulons that are up regulated during growth on glycerol, the ResD regulon is induced, giving a hint that there might be oxygen limitation during growth on glycerol (Fig. 29A). During oxygen limitation, the SigB mediated general stress response is coupled to the transcriptional activation of Fnr regulon and the two-component regulatory systems ResD-ResE (Reents *et al.*, 2006). The cells grow anaerobically either by respiration using nitrate as an electron acceptor and fermentation in the absence of electron acceptors (Nakano & Hulett, 1997). A part of the general starvation regulon controlled by SigH and of the AbrB regulon was induced (Fig. 29E and 29F) as well as a part of the Fur regulon, indicating iron limitation, was induced during growth on glycerol (Fig. 29B). In addition, during growth with glycerol as carbon source the tiling array data revealed the induction of genes encoding spore coat proteins and the *pst* operon encoding an ABC phosphate transporter.

The key genes that are involved in biofilm formation include early stage sporulation genes like *spo0A*, *spo0H* and *abrB*, the *eps* operon comprising 15 genes that are involved in the synthesis of extracellular polysaccharides, and the genes *yhxB*, *yqeK*, *sfp*, *sipW*, *ecsB*, *ylbF* and *ymcA* (Branda *et al.,* 2004). The induction of all these genes during biofilm formation at 36h and 60h could not be seen in high levels, due to the significant induction during the control condition, 24h biofilm formation. From the tiling array analysis, it was evident that the majority of genes of the PhoP regulon regulated during phosphate limitation was induced, especially the *pst* operon encoding an ABC phosphate transporter was 20-fold induced during biofilm formation (Fig. 29C). The two-component regulatory system PhoP-PhoR regulates the gene expression under phosphate limiting conditions to increase the accessibility of alternative phosphate resources. Thus, the induction of the SigB regulon during biofilm formation might be due to the phosphate limitation, one of the classical SigB inducing conditions. Furthermore, the phosphate limitation additionally triggers expression of CodY and SigH regulated genes (Tam et al., 2006). CodY regulates genes in response to the availability of branched chain amino acids and GTP, whereas SigH controls several genes of the early stationary phase process (Tam le *et al.*, 2006). In the tiling array study only a part of the SigH regulon was up regulated under the condition of biofilm

formation. A strong up-regulation of metal limiting regulons, which are involved in the regulation of high affinity uptake systems, and controlled by Fur and MntR (Fig. 29B), respectively, was observed indicating the limiting iron and manganese levels during biofilm formation (Bsat *et al.*, 1998; Gaballa & Helmann, 1998; Que & Helmann, 2000).

Thus, during the conditions like growth on plates, swarming cells, biofilm formation and growth on glycerol, the induction of SigB-mediated general stress response was due to at least one of the stress conditions including oxygen limitation, glucose limitation and phosphate limitation. These insights substantiate that the investigation of multiple regulons under various experimental conditions is necessary to understand the global gene regulation as the genes in the regulons are tightly connected to a complex adaptational network (Bernhardt *et al.*, 2003; Hecker & Völker, 2001).

4.2.4 NEW INSIGHTS INTO THE OSMOTIC REGULATION OF *BACILLUS SUBTILIS*

For the soil-living bacteria like *B. subtilis*, one of the most common stress conditions due to the frequent flooding of the soil is the osmotic stress. In addition to the general stress response that is activated during osmo stress, the cell has a specific regulatory mechanism that induces a set of genes to balance the osmolarity of the cell. A whole-genome transcriptional study, which was conducted by Steil and co-workers on a *sigB* knock out mutant revealed genes that are regulated during salt shock and high osmolar growth (Steil *et al.*, 2003). The induction ratios of the induced osmo-adaptive genes were significantly higher in this study using the *sigB* knockout mutant compared to the tiling array study using the wild type strain. This is in agreement to a similar comparison of the Steil *et al* study to a detailed transcriptional and proteomics study on the wild type strain by Hahne and co-workers, which provides a whole genome profile of osmotic response during salt shock (Hahne *et al.*, 2010). One possible explanation for this higher induction of SigA-regulated osmo-adaptive genes in the *sigB* mutant strain compared to the wild type strain is the sigma factor competition of the RNA polymerase (Rollenhagen *et al.*, 2003). In the wild type strain, during stress conditions, a large number of SigB-dependent genes were rapidly up regulated and a significant proportion of RNA polymerase is engaged in this SigB-dependent transcription. In the phosphate-starved cells, the studies of Antelmann *et al.* and Pragai and Harwood suggested that

SigB is competing with other sigma factors, primarily SigA (Antelmann *et al.*, 2000; Pragai & Harwood, 2002). Thus, in the *sigB* mutant strain, all the RNA polymerase is able to act on the SigA promoters of the osmo-adaptive genes contributing to their higher induction levels.

In total, the tiling array analysis identified 467 differentially regulated newly annotated features, among which 257 are induced and 210 are repressed during salt shock. During continuous growth at high osmolarity, 251 new features were identified, 120 of them were induced and 131 features were repressed. The use of high resolution tiling arrays revealed a significant proportion of antisense RNAs that overlap with transcription of the sense strand. A detailed investigation of the osmo-adaptive genes revealed an antisense RNA targeting the *opuB* operon. The studies of specific sense and antisense pairs demonstrate the effects of such antisense RNAs on their target gene, which can be at the transcriptional level by promoting RNA degradation, effecting RNA stabilization, enhancing transcriptional termination and transcriptional interference (Eiamphungporn & Helmann, 2009; Lasa *et al.*, 2011; Thomason & Storz, 2010). On the other hand, translational effects of antisense RNAs include translational regulation by either blocking or promoting ribosome binding. During the tiling array analysis, the antisense RNAs are more often initiated from non-SigA promoters than protein coding RNAs. Additionally a negative correlation of the expression of sense and antisense pairs appears more often than a co-expression (Nicolas *et al.*, 2012).

The SigB-dependent antisense RNA *S1290* exhibiting a stringent SigB motif was identified in this study, targeting the osmotically inducible *opuB* operon specifically transporting choline into the cell. The induction level of *S1290* is far higher during salt shock compared to the continuous growth at high osmolarity. This pattern is expected for strictly SigB-dependent genes because SigB is only transiently activated following salt shock. Thus, during salt shock condition, the antisense RNA is transcribed in much higher levels compared to its sense RNA (*opuBA-BB-BC-BD*) and might thereby affect the transcription of the osmotically regulated *opuB* operon. Even though the whole operon was targeted, the effect was much more prominent on the *opuBC* gene (Fig. 22). The transcriptional effect of the antisense RNA *S1290* on the *opuB* operon was not observed during continuous growth at high osmolarity. Most probably, the low-level

transcription of the antisense RNA during high osmolar growth compared to its sense strand results in no significant effect on the target gene transcription.

Until today, all the previous reports show that the osmo inducible conditions of salt shock and high osmolar growth induce all the opu genes. In addition to the non-induction of *opuB* operon during salt shock, we have also observed that the *opuC* operon was not induced during continuous growth at high osmolarity. It can be assumed that when the cells are adapted to the high osmotic conditions the induction of the *opuC* operon, which is an ATP transporter with a broad range of substrates, is not quite necessary to the cell. However, the underlying reason for the non-induction of *opuC* operon is not yet completely known. The previous array studies that were conducted to analyze the osmotic response in *B. subtilis* used PCR products or oligonucleotide probes. Due to the high sequence similarity of *opuCA* and *opuBA* (78%), there could have been cross hybridization artifacts and thereby the resultant induction levels in the previous studies might be false positive.

4.2.5 SPECIFIC EXPRESSION PATTERNS OF OSMO-ADAPTIVE GENES IN CERTAIN CONDITIONS

In addition to the osmo inducible conditions of salt shock and continuous growth, the osmo-adaptive genes exhibit significant lower expression during growth in LB medium compared to minimal medium. The lower expression is reproducible in a second completely independent dataset which was generated using the Agilent platform (Ulrike Mäder, personal communication). Even though the underlying reasons are not known, the strong repression of osmo-adaptive genes in LB medium, which is widely used in the industrial applications, is of considerable interest in further studies. The repression of osmo-adaptive genes during diamide stress might be due to the indirect effect via a protein (regulator) which is effected by S-S bridge formation (Fig. 26).

The gene expression of the dual regulated genes *opuD* and *opuE*, which are additionally controlled by the alternative sigma factor, SigB, showed an unexpected pattern during stationary phase growth. When the cells were grown in minimal medium, the expression of *opuD* was slightly increased whereas the expression of *opuE* remained unchanged (Fig. 23). In a detailed time series experiment using a *sigK* mutant strain, performed in modified M9 medium following the glucose starvation process, the expression profiles of the *opuD* and *opuE* genes behave similar (Fig. 27). Even though both genes possess stringent SigB motifs, for yet unknown reasons no induction was observed during stationary phase.

The *proHJ* operon, which is under the control of early mother-cell specific sporulation sigma factor SigE, was transiently induced during sporulation conditions (Fig. 25A). The expression starts at 2h after the onset of sporulation and is significantly higher until 5h after initiation of sporulation. The tiling array analysis approved the presence of a stringent SigE promoter upstream of the operon. In a previous study from Steil and co-workers the *proHJ* operon was also assigned to the SigE regulon due to the expression during the sporulation process (Steil *et al.*, 2005). The genes that encode the spore coat components are under the control of SigE and a high proportion of the proteins synthesized during sporulation are part of this spore coat. The percentage of proline within the spore coat proteins was slightly, but significantly higher compared to all the

non-spore proteins (Fig. 25b). This might indicate that proline and the alternative proline biosynthesis might play a role during the generation of resistant spores.

In conclusion, the tiling array analysis revealed a set of newly annotated RNA features under the control of the alternative sigma factor, SigB, as well as during the osmotic response. This study focuses on a subgroup of 32 genes revealing the SigB motif with significant low induction profiles in all the tested SigB-inducing stress conditions, which potentially possess secondary regulators and are highly relevant for further studies. The current work further identifies new conditions like growth on plates, swarming cells, biofilm formation and growth on glycerol as carbon source, wherein the induction of SigB-dependent genes was significant. The RNA new features, the new SigB-inducing conditions and the specific conditions that target the osmo specific genes will serve as inputs for further studies.

5. REFERENCES

Akbar, S., Kang, C. M., Gaidenko, T. A. & Price, C. W. (1997). Modulator protein RsbR regulates environmental signalling in the general stress pathway of *Bacillus subtilis*. *Mol Microbiol* **24**, 567-578.

Alper, S., Dufour, A., Garsin, D. A., Duncan, L. & Losick, R. (1996). Role of adenosine nucleotides in the regulation of a stress-response transcription factor in *Bacillus subtilis*. *J Mol Biol* **260**, 165-177.

Anagnostopoulos, C. & Spizizen, J. (1961). Requirements for Transformation in *Bacillus Subtilis*. *J Bacteriol* **81**, 741-746.

Antelmann, H., Engelmann, S., Schmid, R. & Hecker, M. (1996). General and oxidative stress responses in *Bacillus subtilis*: cloning, expression, and mutation of the alkyl hydroperoxide reductase operon. *J Bacteriol* **178**, 6571-6578.

Antelmann, H., Scharf, C. & Hecker, M. (2000). Phosphate starvation-inducible proteins of *Bacillus subtilis*: proteomics and transcriptional analysis. *J Bacteriol* **182**, 4478-4490.

Au, N., Kuester-Schoeck, E., Mandava, V. & other authors (2005). Genetic composition of the *Bacillus subtilis* SOS system. *J Bacteriol* **187**, 7655-7666.

Avila-Pérez, M., Hellingwerf, K. J. & Kort, R. (2006). Blue light activates the sigmaB-dependent stress response of *Bacillus subtilis* via YtvA. *J Bacteriol* **188**, 6411-6414.

Avila-Pérez, M., van der Steen, J. B., Kort, R. & Hellingwerf, K. J. (2010). Red light activates the sigmaB-mediated general stress response of *Bacillus subtilis* via the energy branch of the upstream signaling cascade. *J Bacteriol* **192**, 755-762.

Banse, A. V., Chastanet, A., Rahn-Lee, L., Hobbs, E. C. & Losick, R. (2008). Parallel pathways of repression and antirepression governing the transition to stationary phase in *Bacillus subtilis*. *Proc Natl Acad Sci U S A* **105**, 15547-15552.

Baranova, N. N., Danchin, A. & Neyfakh, A. A. (1999). Mta, a global MerR-type regulator of the *Bacillus subtilis* multidrug-efflux transporters. *Mol Microbiol* **31**, 1549-1559.

Barbe, V., Cruveiller, S., Kunst, F. & other authors (2009). From a consortium sequence to a unified sequence: the *Bacillus subtilis* 168 reference genome a decade later. *Microbiology* **155**, 1758-1775.

Barrett, J. H. & Cairns, D. A. (2008). Application of the random forest classification method to peaks detected from mass spectrometric proteomic profiles of cancer patients and controls. *Stat Appl Genet Mol Biol* **7**, Article4.

Benson, A. K. & Haldenwang, W. G. (1993). *Bacillus subtilis* sigma B is regulated by a binding protein (RsbW) that blocks its association with core RNA polymerase. *Proc Natl Acad Sci U S A* **90**, 2330-2334.

Bernhardt, J., Weibezahn, J., Scharf, C. & Hecker, M. (2003). *Bacillus subtilis* during feast and famine: visualization of the overall regulation of protein synthesis during glucose starvation by proteome analysis. *Genome Res* **13**, 224-237.

Boch, J., Kempf, B. & Bremer, E. (1994). Osmoregulation in *Bacillus subtilis*: synthesis of the osmoprotectant glycine betaine from exogenously provided choline. *J Bacteriol* **176**, 5364-5371.

Bower, S., Perkins, J. B., Yocum, R. R., Howitt, C. L., Rahaim, P. & Pero, J. (1996). Cloning, sequencing, and characterization of the *Bacillus subtilis* biotin biosynthetic operon. *J Bacteriol* **178**, 4122-4130.

Boylan, S. A., Thomas, M. D. & Price, C. W. (1991). Genetic method to identify regulons controlled by nonessential elements: isolation of a gene dependent on alternate transcription factor sigma B of *Bacillus subtilis. J Bacteriol* **173**, 7856-7866.

Boylan, S. A., Rutherford, A., Thomas, S. M. & Price, C. W. (1992). Activation of *Bacillus subtilis* transcription factor sigma B by a regulatory pathway responsive to stationary-phase signals. *J Bacteriol* **174**, 3695-3706.

Boylan, S. A., Redfield, A. R., Brody, M. S. & Price, C. W. (1993). Stress-induced activation of the sigma B transcription factor of *Bacillus subtilis. J Bacteriol* **175**, 7931-7937.

Breiman, L. (2001). Random forests. *Machine Learning* **45**, 5-32.

Bremer, E. (2002). Adaptation to Changing Osmolarity. In *Bacillus subtilis and Its Closest Relatives From Genes to Cells*, pp. 385-391. Edited by A. L. Sonnenshein, J. A. Hoch, R. Losick & American Society for Microbiology. Washington, DC.: ASM Press.

Brigulla, M., Hoffmann, T., Krisp, A., Völker, A., Bremer, E. & Völker, U. (2003). Chill induction of the SigB-dependent general stress response in *Bacillus subtilis* and its contribution to low-temperature adaptation. *J Bacteriol* **185**, 4305-4314.

Bsat, N., Herbig, A., Casillas-Martinez, L., Setlow, P. & Helmann, J. D. (1998). *Bacillus subtilis* contains multiple Fur homologues: identification of the iron uptake (Fur) and peroxide regulon (PerR) repressors. *Mol Microbiol* **29**, 189-198.

Burbulys, D., Trach, K. A. & Hoch, J. A. (1991). Initiation of sporulation in *B. subtilis* is controlled by a multicomponent phosphorelay. *Cell* **64**, 545-552.

Burkholder, P. R. & Giles, N. H., Jr. (1947). Induced biochemical mutations in *Bacillus subtilis. Am J Bot* **34**, 345-348.

Cao, M., Kobel, P. A., Morshedi, M. M., Wu, M. F., Paddon, C. & Helmann, J. D. (2002). Defining the *Bacillus subtilis* sigma(W) regulon: a comparative analysis of promoter consensus search, run-off transcription/macroarray analysis (ROMA), and transcriptional profiling approaches. *J Mol Biol* **316**, 443-457.

Chen, L., Keramati, L. & Helmann, J. D. (1995). Coordinate regulation of *Bacillus subtilis* peroxide stress genes by hydrogen peroxide and metal ions. *Proc Natl Acad Sci U S A* **92**, 8190-8194.

Chen, X. W. & Liu, M. (2005). Prediction of protein-protein interactions using random decision forest framework. *Bioinformatics* **21**, 4394-4400.

Chu, F., Kearns, D. B., Branda, S. S., Kolter, R. & Losick, R. (2006). Targets of the master regulator of biofilm formation in *Bacillus subtilis*. *Mol Microbiol* **59**, 1216-1228.

Chumsakul, O., Takahashi, H., Oshima, T., Hishimoto, T., Kanaya, S., Ogasawara, N. & Ishikawa, S. (2011). Genome-wide binding profiles of the *Bacillus subtilis* transition state regulator AbrB and its homolog Abh reveals their interactive role in transcriptional regulation. *Nucleic Acids Res* **39**, 414-428.

Comella, N. & Grossman, A. D. (2005). Conservation of genes and processes controlled by the quorum response in bacteria: characterization of genes controlled by the quorum-sensing transcription factor ComA in *Bacillus subtilis*. *Mol Microbiol* **57**, 1159-1174.

Csonka, L. N. (1989). Physiological and genetic responses of bacteria to osmotic stress. *Microbiol Rev* **53**, 121-147.

Datta, S. (2008). Classification of breast cancer versus normal samples from mass spectrometry profiles using linear discriminant analysis of important features selected by random forest. *Stat Appl Genet Mol Biol* **7**, Article7.

Demsar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *J Mach Learn Res* **7**, 1-30.

Derre, I., Rapoport, G. & Msadek, T. (1999). CtsR, a novel regulator of stress and heat shock response, controls *clp* and molecular chaperone gene expression in gram-positive bacteria. *Mol Microbiol* **31**, 117-131.

Diaz-Uriarte, R. & Alvarez de Andres, S. (2006). Gene selection and classification of microarray data using random forest. *BMC Bioinformatics* **7**, 3.

Drzewiecki, K., Eymann, C., Mittenhuber, G. & Hecker, M. (1998). The *yvyD* gene of *Bacillus subtilis* is under dual control of sigmaB and sigmaH. *J Bacteriol* **180**, 6674-6680.

Dufour, A. & Haldenwang, W. G. (1994). Interactions between a *Bacillus subtilis* anti-sigma factor (RsbW) and its antagonist (RsbV). *J Bacteriol* **176**, 1813-1820.

Ebbole, D. J. & Zalkin, H. (1987). Cloning and characterization of a 12-gene cluster from *Bacillus subtilis* encoding nine enzymes for de novo purine nucleotide synthesis. *J Biol Chem* **262**, 8274-8287.

Efron, B. (1982). *The jackknife, the bootstrap, and other resampling plans*. Philadelphia, Pa.: Society for Industrial and Applied Mathematics.

Ehrenberg, C. G. (1835). Dritter Beitrag zur Erkenntniss grosser Organization in der Richtung des kleinsten Raumes. *Physikal Abh Königlich Akad Wiss Berl Jahr* **1833–1835**, 135–336.

Eiamphungporn, W. & Helmann, J. D. (2008). The *Bacillus subtilis* sigma(M) regulon and its contribution to cell envelope stress responses. *Mol Microbiol* **67**, 830-848.

Eiamphungporn, W. & Helmann, J. D. (2009). Extracytoplasmic function sigma factors regulate expression of the *Bacillus subtilis yabE* gene via a cis-acting antisense RNA. *J Bacteriol* **191**, 1101-1105.

Errington, J. (2003). Regulation of endospore formation in *Bacillus subtilis. Nat Rev Microbiol* **1**, 117-126.

Erwin, K. N., Nakano, S. & Zuber, P. (2005). Sulfate-dependent repression of genes that function in organosulfur metabolism in *Bacillus subtilis* requires Spx. *J Bacteriol* **187**, 4042-4049.

Eymann, C., Mach, H., Harwood, C. R. & Hecker, M. (1996). Phosphate-starvation-inducible proteins in *Bacillus subtilis*: a two-dimensional gel electrophoresis study. *Microbiology* **142 (Pt 11)**, 3163-3170.

Florez, L. A., Roppel, S. F., Schmeisky, A. G., Lammers, C. R. & Stülke, J. (2009). A community-curated consensual annotation that is continuously updated: the *Bacillus subtilis* centred wiki SubtiWiki. *Database (Oxford)* **2009**, bap012.

Fujita, Y. (2009). Carbon catabolite control of the metabolic network in *Bacillus subtilis. Biosci Biotechnol Biochem* **73**, 245-259.

Gaballa, A. & Helmann, J. D. (1998). Identification of a zinc-specific metalloregulatory protein, Zur, controlling zinc transport operons in *Bacillus subtilis. J Bacteriol* **180**, 5815-5821.

Gaidenko, T. A. & Price, C. W. (1998). General stress transcription factor sigmaB and sporulation transcription factor sigmaH each contribute to survival of *Bacillus subtilis* under extreme growth conditions. *J Bacteriol* **180**, 3730-3733.

Grossman, A. D. (1991). Integration of developmental signals and the initiation of sporulation in *B. subtilis. Cell* **65**, 5-8.

Grundy, F. J. & Henkin, T. M. (1998). The S box regulon: a new global transcription termination control system for methionine and cysteine biosynthesis genes in gram-positive bacteria. *Mol Microbiol* **30**, 737-749.

Hahne, H., Mäder, U., Otto, A., Bonn, F., Steil, L., Bremer, E., Hecker, M. & Becher, D. (2010). A comprehensive proteomics and transcriptomics analysis of *Bacillus subtilis* salt stress adaptation. *J Bacteriol* **192**, 870-882.

Haldenwang, W. G. & Losick, R. (1980). Novel RNA polymerase sigma factor from *Bacillus subtilis. Proc Natl Acad Sci U S A* **77**, 7000-7004.

Hecker, M. & Völker, U. (2001). General stress response of *Bacillus subtilis* and other bacteria. *Adv Microb Physiol* **44**, 35-91.

Hecker, M., Pane-Farre, J. & Völker, U. (2007). SigB-dependent general stress response in *Bacillus subtilis* and related gram-positive bacteria. *Annu Rev Microbiol* **61**, 215-236.

Helmann, J. D., Wu, M. F., Kobel, P. A., Gamo, F. J., Wilson, M., Morshedi, M. M., Navre, M. & Paddon, C. (2001). Global transcriptional response of *Bacillus subtilis* to heat shock. *J Bacteriol* **183**, 7318-7328.

Helmann, J. D., Wu, M. F., Gaballa, A., Kobel, P. A., Morshedi, M. M., Fawcett, P. & Paddon, C. (2003). The global transcriptional response of *Bacillus subtilis* to peroxide stress is coordinated by three transcription factors. *J Bacteriol* **185**, 243-253.

Henkin, T. M. (1996). The role of CcpA transcriptional regulator in carbon metabolism in *Bacillus subtilis. FEMS Microbiol Lett* **135**, 9-15.

Hertwig, F. (2003). Analyse der Struktur und der Kinetik der Induktion des SigB-Regulons von *Bacillus subtilis* mit Hilfe von DNA-Microarrays. In *Fachbereich Biologie.* Marburg: Philipps-Universität Marburg.

Higgins, D. & Dworkin, J. (2012). Recent progress in *Bacillus subtilis* sporulation. *FEMS Microbiol Rev* **36**, 131-148.

Hilbert, D. W. & Piggot, P. J. (2004). Compartmentalization of gene expression during *Bacillus subtilis* spore formation. *Microbiol Mol Biol Rev* **68**, 234-262.

Hoffmann, T. & Bremer, E. (2011). Protection of *Bacillus subtilis* against cold stress via compatible-solute acquisition. *J Bacteriol* **193**, 1552-1562.

Höper, D., Völker, U. & Hecker, M. (2005). Comprehensive characterization of the contribution of individual SigB-dependent general stress genes to stress resistance of *Bacillus subtilis. J Bacteriol* **187**, 2810-2826.

Hulett, F. M. (2002). The Pho regulon. In *Bacillus subtilis and Its Closest Relatives From Genes to Cells*, pp. 193–203. Edited by A. L. Sonnenshein, J. A. Hoch, R. Losick & American Society for Microbiology. Washington, DC.: ASM Press.

Igoshin, O. A., Brody, M. S., Price, C. W. & Savageau, M. A. (2007). Distinctive topologies of partner-switching signaling networks correlate with their physiological roles. *J Mol Biol* **369**, 1333-1352.

Irnov, I., Sharma, C. M., Vogel, J. & Winkler, W. C. (2010). Identification of regulatory RNAs in *Bacillus subtilis. Nucleic Acids Res* **38**, 6637-6651.

Izmirlian, G. (2004). Application of the random forest classification algorithm to a SELDI-TOF proteomics study in the setting of a cancer prevention trial. *Ann N Y Acad Sci* **1020**, 154-174.

Jervis, A. J., Thackray, P. D., Houston, C. W., Horsburgh, M. J. & Moir, A. (2007). SigM-responsive genes of *Bacillus subtilis* and their promoters. *J Bacteriol* **189**, 4534-4538.

Jürgen, B., Tobisch, S., Wumpelmann, M., Gordes, D., Koch, A., Thurow, K., Albrecht, D., Hecker, M. & Schweder, T. (2005). Global expression profiling of *Bacillus subtilis* cells during industrial-close fed-batch fermentations with different nitrogen sources. *Biotechnol Bioeng* **92**, 277-298.

Kang, C. M., Brody, M. S., Akbar, S., Yang, X. & Price, C. W. (1996). Homologous pairs of regulatory proteins control activity of *Bacillus subtilis* transcription factor sigma(b) in response to environmental stress. *J Bacteriol* **178**, 3846-3853.

Kearns, D. B., Chu, F., Branda, S. S., Kolter, R. & Losick, R. (2005). A master regulator for biofilm formation by *Bacillus subtilis. Mol Microbiol* **55**, 739-749.

Klein-Seetharaman, J., Tastan, O., Qi, Y. J. & Carbonell, J. G. (2009). Prediction of Interactions between Hiv-1 and Human Proteins by Information Integration. *Pacific Symposium on Biocomputing 2009*, 516-527.

Kok, J., Trach, K. A. & Hoch, J. A. (1994). Effects on *Bacillus subtilis* of a conditional lethal mutation in the essential GTP-binding protein Obg. *J Bacteriol* **176**, 7155-7160.

Krüger, E., Msadek, T. & Hecker, M. (1996). Alternate promoters direct stress-induced transcription of the *Bacillus subtilis clpC* operon. *Mol Microbiol* **20**, 713-723.

Kumaraswami, M., Newberry, K. J. & Brennan, R. G. (2010). Conformational plasticity of the coiled-coil domain of BmrR is required for bmr operator binding: the structure of unliganded BmrR. *J Mol Biol* **398**, 264-275.

Kunst, F., Ogasawara, N., Moszer, I. & other authors (1997). The complete genome sequence of the gram-positive bacterium *Bacillus subtilis. Nature* **390**, 249-256.

Lasa, I., Toledo-Arana, A., Dobin, A. & other authors (2011). Genome-wide antisense transcription drives mRNA processing in bacteria. *Proc Natl Acad Sci U S A* **108**, 20172-20177.

Leelakriangsak, M., Kobayashi, K. & Zuber, P. (2007). Dual negative control of spx transcription initiation from the P3 promoter by repressors PerR and YodB in *Bacillus subtilis. J Bacteriol* **189**, 1736-1744.

Lei, J., Zhou, Y. F., Li, L. F. & Su, X. D. (2009). Structural and biochemical analyses of YvgN and YtbE from *Bacillus subtilis. Protein Sci* **18**, 1792-1800.

Losick, R. & Pero, J. (1981). Cascades of Sigma factors. *Cell* **25**, 582-584.

Losick, R., Youngman, P. & Piggot, P. J. (1986). Genetics of endospore formation in *Bacillus subtilis. Annu Rev Genet* **20**, 625-669.

Mäder, U., Homuth, G., Völker, U. & Hecker, M. (2007). Global regulatory systems mediating general starvation responses in *Bacillus subtilis*. In *Global Regulatory Networks in Bacillus subtilis*. Edited by Y. Fujita: Transworld Research Network.

Mäder, U., Schmeisky, A. G., Flórez, L. A. & Stülke, J. (2012). SubtiWiki--a comprehensive community resource for the model organism *Bacillus subtilis. Nucleic Acids Res* **40**, D1278-1287.

Manilich, E. A., Ozsoyoglu, Z. M., Trubachev, V. & Radivoyevitch, T. (2011). Classification of large microarray datasets using fast random forest construction. *J Bioinform Comput Biol* **9**, 251-267.

Marvasi, M., Visscher, P. T. & Casillas Martinez, L. (2010). Exopolymeric substances (EPS) from *Bacillus subtilis*: polymers and genes encoding their synthesis. *FEMS Microbiol Lett* **313**, 1-9.

Mendez, M. B., Orsaria, L. M., Philippe, V., Pedrido, M. E. & Grau, R. R. (2004). Novel roles of the master transcription factors Spo0A and sigmaB for survival and sporulation of *Bacillus subtilis* at low growth temperature. *J Bacteriol* **186**, 989-1000.

Moorthy, K. & Mohamad, M. S. (2011). Random forest for gene selection and microarray data classification. *Bioinformation* **7**, 142-146.

Mostertz, J., Scharf, C., Hecker, M. & Homuth, G. (2004). Transcriptome and proteome analysis of *Bacillus subtilis* gene expression in response to superoxide and peroxide stress. *Microbiology* **150**, 497-512.

Nakano, M. M. & Hulett, F. M. (1997). Adaptation of *Bacillus subtilis* to oxygen limitation. *FEMS Microbiol Lett* **157**, 1-7.

Nakano, S., Kuster-Schock, E., Grossman, A. D. & Zuber, P. (2003). Spx-dependent global transcriptional control is induced by thiol-specific oxidative stress in *Bacillus subtilis. Proc Natl Acad Sci U S A* **100**, 13603-13608.

Nannapaneni, P., Hertwig, F., Depke, M., Hecker, M., Mäder, U., Völker, U., Steil, L. & van Hijum, S. A. (2012). Defining the structure of the general stress regulon of *Bacillus*

subtilis using targeted microarray analysis and random forest classification. *Microbiology* **158**, 696-707.

Nguyen, T. T., Eiamphungporn, W., Mäder, U., Liebeke, M., Lalk, M., Hecker, M., Helmann, J. D. & Antelmann, H. (2009). Genome-wide responses to carbonyl electrophiles in *Bacillus subtilis*: control of the thiol-dependent formaldehyde dehydrogenase *A*dhA and cysteine proteinase YraA by the MerR-family regulator YraB (AdhR). *Mol Microbiol* **71**, 876-894.

Nicolas, P., Leduc, A., Robin, S., Rasmussen, S., Jarmer, H. & Bessières, P. (2009). Transcriptional landscape estimation from tiling array data using a model of signal shift and drift. *Bioinformatics* **25**, 2341-2347.

Nicolas, P., Mäder, U., Dervyn, E. & other authors (2012). Condition-dependent transcriptome reveals high-level regulatory architecture in *Bacillus subtilis*. *Science* **335**, 1103-1106.

Perkins, J. B., Bower, S., Howitt, C. L., Yocum, R. R. & Pero, J. (1996). Identification and characterization of transcripts from the biotin biosynthetic operon of *Bacillus subtilis*. *J Bacteriol* **178**, 6361-6365.

Petersohn, A., Antelmann, H., Gerth, U. & Hecker, M. (1999a). Identification and transcriptional analysis of new members of the sigmaB regulon in *Bacillus subtilis*. *Microbiology* **145 (Pt 4)**, 869-880.

Petersohn, A., Bernhardt, J., Gerth, U., Höper, D., Koburger, T., Völker, U. & Hecker, M. (1999b). Identification of sigma(B)-dependent genes in *Bacillus subtilis* using a promoter consensus-directed search and oligonucleotide hybridization. *J Bacteriol* **181**, 5718-5724.

Petersohn, A., Engelmann, S., Setlow, P. & Hecker, M. (1999c). The *katX* gene of *Bacillus subtilis* is under dual control of sigmaB and sigmaF. *Mol Gen Genet* **262**, 173-179.

Petersohn, A., Brigulla, M., Haas, S., Hoheisel, J. D., Völker, U. & Hecker, M. (2001). Global analysis of the general stress response of *Bacillus subtilis*. *J Bacteriol* **183**, 5617-5631.

Piggot, P. J. & Losick, R. (2002). Sporulation genes and intercomparmental regulation. In *Bacillus subtilis and Its Closest Relatives From Genes to Cells*, pp. 483–518. Edited by A. L. Sonnenshein, J. A. Hoch, R. Losick & American Society for Microbiology. Washington, DC.: ASM Press.

Potts, M. (1994). Desiccation tolerance of prokaryotes. *Microbiol Rev* **58**, 755-805.

Pragai, Z. & Harwood, C. R. (2002). Regulatory interactions between the Pho and sigma(B)-dependent general stress regulons of *Bacillus subtilis*. *Microbiol-Sgm* **148**, 1593-1602.

Price, C. W. (2000). Protective function and regulation of the general stress response in *Bacillus subtilis* and related gram-positive bacteria. In *Bacterial stress responses*, pp. 179-197. Edited by G. Storz, R. Hengge & American Society for Microbiology.: ASM Press, Washington, D.C.

Price, C. W., Fawcett, P., Ceremonie, H., Su, N., Murphy, C. K. & Youngman, P. (2001). Genome-wide analysis of the general stress response in *Bacillus subtilis*. *Mol Microbiol* **41**, 757-774.

Price, C. W. (2002). General stress response. In *Bacillus subtilis and Its Closest Relatives From Genes to Cells*, pp. 369–384. Edited by A. L. Sonnenshein, J. A. Hoch, R. Losick & American Society for Microbiology. Washington, DC.: ASM Press.

Price, C. W. (2011). General Stress Response in *Bacillus subtilis* and Related Gram-Positive Bacteria. In *Bacterial stress responses*. Edited by G. Storz, R. Hengge & American Society for Microbiology. Washington, DC: ASM Press.

Que, Q. & Helmann, J. D. (2000). Manganese homeostasis in *Bacillus subtilis* is regulated by MntR, a bifunctional regulator related to the diphtheria toxin repressor family of proteins. *Mol Microbiol* **35**, 1454-1468.

Rasmussen, S., Nielsen, H. B. & Jarmer, H. (2009). The transcriptionally active regions in the genome of *Bacillus subtilis*. *Mol Microbiol* **73**, 1043-1057.

Reder, A., Höper, D., Weinberg, C., Gerth, U., Fraunholz, M. & Hecker, M. (2008). The Spx paralogue MgsR (YqgZ) controls a subregulon within the general stress response of *Bacillus subtilis*. *Mol Microbiol* **69**, 1104-1120.

Reder, A., Albrecht, D., Gerth, U. & Hecker, M. (2012a). Cross-talk between the general stress response and sporulation initiation in *Bacillus subtilis* - the sigma(B) promoter of *spo0E* represents an AND-gate. *Environ Microbiol*.

Reder, A., Gerth, U. & Hecker, M. (2012b). Integration of sigmaB activity into the decision-making process of sporulation initiation in *Bacillus subtilis*. *J Bacteriol* **194**, 1065-1074.

Reder, A., Höper, D., Gerth, U. & Hecker, M. (2012c). The contribution of individual sigmaB-dependent general stress genes to oxidative stress resistance of *Bacillus subtilis*. *J Bacteriol*.

Reents, H., Munch, R., Dammeyer, T., Jahn, D. & Hartig, E. (2006). The Fnr regulon of *Bacillus subtilis*. *J Bacteriol* **188**, 1103-1112.

Rollenhagen, C., Antelmann, H., Kirstein, J., Delumeau, O., Hecker, M. & Yudkin, M. D. (2003). Binding of sigma(A) and sigma(B) to core RNA polymerase after environmental stress in *Bacillus subtilis*. *J Bacteriol* **185**, 35-40.

Scharf, C., Riethdorf, S., Ernst, H., Engelmann, S., Völker, U. & Hecker, M. (1998). Thioredoxin is an essential protein induced by multiple stresses in *Bacillus subtilis. J Bacteriol* **180**, 1869-1877.

Scott, J. M. & Haldenwang, W. G. (1999). Obg, an essential GTP binding protein of *Bacillus subtilis*, is necessary for stress activation of transcription factor sigma(B). *J Bacteriol* **181**, 4653-4660.

Sekowska, A. & Danchin, A. (2002). The methionine salvage pathway in *Bacillus subtilis. BMC Microbiol* **2**, 8.

Shi, T., Seligson, D., Belldegrun, A. S., Palotie, A. & Horvath, S. (2005). Tumor classification by tissue microarray profiling: random forest clustering applied to renal cell carcinoma. *Mod Pathol* **18**, 547-557.

Sierro, N., Makita, Y., de Hoon, M. & Nakai, K. (2008). DBTBS: a database of transcriptional regulation in *Bacillus subtilis* containing upstream intergenic conservation information. *Nucleic Acids Res* **36**, D93-96.

Singh, K. D., Schmalisch, M. H., Stülke, J. & Gorke, B. (2008). Carbon catabolite repression in *Bacillus subtilis*: quantitative analysis of repression exerted by different carbon sources. *J Bacteriol* **190**, 7275-7284.

Staden, R. (1984). Computer methods to locate signals in nucleic acid sequences. *Nucleic Acids Res* **12**, 505-519.

Steil, L., Hoffmann, T., Budde, I., Völker, U. & Bremer, E. (2003). Genome-wide transcriptional profiling analysis of adaptation of *Bacillus subtilis* to high salinity. *J Bacteriol* **185**, 6358-6370.

Steil, L. (2004). Charakterisierung der Anpassung von *Bacillus subtilis* an verschiedene Unweltstresse mittels Transcriptom- und Proteomanalysen. In *Mathematisch-Naturwissenschaftichen Fakultät*. Greifswald: Ernst-Moritz-Arndt-Universität Greifswald.

Steil, L., Serrano, M., Henriques, A. O. & Völker, U. (2005). Genome-wide analysis of temporally regulated and compartment-specific gene expression in sporulating cells of *Bacillus subtilis. Microbiology* **151**, 399-420.

Stragier, P. & Losick, R. (1996). Molecular genetics of sporulation in *Bacillus subtilis. Annu Rev Genet* **30**, 297-241.

Strobl, C., Boulesteix, A. L., Zeileis, A. & Hothorn, T. (2007). Bias in random forest variable importance measures: illustrations, sources and a solution. *BMC Bioinformatics* **8**, 25.

Tam le, T., Antelmann, H., Eymann, C., Albrecht, D., Bernhardt, J. & Hecker, M. (2006). Proteome signatures for stress and starvation in *Bacillus subtilis* as revealed by a 2-D gel image color coding approach. *Proteomics* **6**, 4565-4585.

Thomason, M. K. & Storz, G. (2010). Bacterial antisense RNAs: how many are there, and what are they doing? *Annu Rev Genet* **44**, 167-188.

Tobisch, S., Zühlke, D., Bernhardt, J., Stülke, J. & Hecker, M. (1999). Role of CcpA in regulation of the central pathways of carbon catabolism in *Bacillus subtilis. J Bacteriol* **181**, 6996-7004.

Trach, K. & Hoch, J. A. (1989). The *Bacillus subtilis spo0B* stage 0 sporulation operon encodes an essential GTP-binding protein. *J Bacteriol* **171**, 1362-1371.

van Hijum, S. A., Garcia de la Nava, J., Trelles, O., Kok, J. & Kuipers, O. P. (2003). MicroPreP: a cDNA microarray data pre-processing framework. *Appl Bioinformatics* **2**, 241-244.

Varon, D., Brody, M. S. & Price, C. W. (1996). *Bacillus subtilis* operon under the dual control of the general stress transcription factor sigma B and the sporulation transcription factor sigma H. *Mol Microbiol* **20**, 339-350.

Vidwans, S. J., Ireton, K. & Grossman, A. D. (1995). Possible role for the essential GTP-binding protein Obg in regulating the initiation of sporulation in *Bacillus subtilis. J Bacteriol* **177**, 3308-3311.

Vijay, K., Brody, M. S., Fredlund, E. & Price, C. W. (2000). A PP2C phosphatase containing a PAS domain is required to convey signals of energy stress to the sigmaB transcription factor of *Bacillus subtilis. Mol Microbiol* **35**, 180-188.

Völker, U., Engelmann, S., Maul, B., Riethdorf, S., Völker, A., Schmid, R., Mach, H. & Hecker, M. (1994). Analysis of the induction of general stress proteins of *Bacillus subtilis. Microbiology* **140 (Pt 4)**, 741-752.

Völker, U., Völker, A., Maul, B., Hecker, M., Dufour, A. & Haldenwang, W. G. (1995). Separate mechanisms activate sigma B of *Bacillus subtilis* in response to environmental and metabolic stresses. *J Bacteriol* **177**, 3771-3780.

Völker, U., Maul, B. & Hecker, M. (1999). Expression of the sigmaB-dependent general stress regulon confers multiple stress resistance in *Bacillus subtilis. J Bacteriol* **181**, 3942-3948.

Waldmüller, S., Freund, P., Mauch, S., Toder, R. & Vosberg, H. P. (2002). Low-density DNA microarrays are versatile tools to screen for known mutations in hypertrophic cardiomyopathy. *Hum Mutat* **19**, 560-569.

Wang, S. T., Setlow, B., Conlon, E. M., Lyon, J. L., Imamura, D., Sato, T., Setlow, P., Losick, R. & Eichenberger, P. (2006). The forespore line of gene expression in *Bacillus subtilis. J Mol Biol* **358**, 16-37.

Weng, M., Nagy, P. L. & Zalkin, H. (1995). Identification of the *Bacillus subtilis* pur operon repressor. *Proc Natl Acad Sci U S A* **92**, 7455-7459.

Whatmore, A. M., Chudek, J. A. & Reed, R. H. (1990). The effects of osmotic upshock on the intracellular solute pools of *Bacillus subtilis*. *J Gen Microbiol* **136**, 2527-2535.

Whatmore, A. M. & Reed, R. H. (1990). Determination of turgor pressure in *Bacillus subtilis*: a possible role for K+ in turgor regulation. *J Gen Microbiol* **136**, 2521-2526.

Wu, X. Y., Wu, Z. Y. & Li, K. (2008). Identification of differential gene expression for microarray data using recursive random forest. *Chin Med J (Engl)* **121**, 2492-2496.

Yang, X., Kang, C. M., Brody, M. S. & Price, C. W. (1996). Opposing pairs of serine protein kinases and phosphatases transmit signals of environmental stress to activate a bacterial transcription factor. *Genes Dev* **10**, 2265-2275.

Yang, Z. R. (2009). Predicting sulfotyrosine sites using the random forest algorithm with significantly improved prediction accuracy. *BMC Bioinformatics* **10**, 361.

You, C., Sekowska, A., Francetic, O., Martin-Verstraete, I., Wang, Y. & Danchin, A. (2008). Spx mediates oxidative stress regulation of the methionine sulfoxide reductases operon in *Bacillus subtilis*. *BMC Microbiol* **8**, 128.

Yuan, G. & Wong, S. L. (1995). Regulation of *groE* expression in *Bacillus subtilis*: the involvement of the sigma A-like promoter and the roles of the inverted repeat sequence (CIRCE). *J Bacteriol* **177**, 5427-5433.

ERKLÄRUNG

Hiermit erkläre ich, dass diese Arbeit bisher von mir weder an der Mathematisch-Naturwissenschaftlichen Fakultät der Ernst-Moritz-Arndt-Universität Greifswald noch einer anderen wissenschaftlichen Einrichtung zum Zwecke der Promotion eingereicht wurde.

Ferner erkläre ich, dass diese Arbeit selbständig verfasst und keine anderen als die darin angegebenen Hilfsmittel benutzt und keine Textabschnitte eines Dritten ohne Kennzeichnung übernommen habe.

Greifswald, August 2012

Priyanka Nannapaneni

CURRICULUM VITAE

Personal Details

Name   Priyanka Nannapaneni

Date of Birth   15.08.1985

Place of Birth   Kodali

Marital Status   Married

Nationality   Indian

Educational Qualifications

Secondary School (1989-2000): Nagarjuna Public School, India

Board of Intermediate Education (2000-2002): Nalanda Girls Junior College, India

Bachelor in Biotechnology (2002-2006): Godavari Institute of Engineering and Technology, Rajahmundry, affiliated to JNTU Hyderabad, India

Masters in Bioinformatics (2006-2008): University of Hyderabad, Hyderabad, India

Doctoral Study (since 08/2008): Doctoral student at Interfakultary Institute for Genetics and Functional Genomics, Dept. of Functional Genomics, Ernst-Moritz-ArndtUniversity, Greifswald, Germany

Greifswald, Date   Signature

Priyanka Nannapaneni, Falk Hertwig, Maren Depke, Michael Hecker, Ulrike Mäder, Uwe Völker, Leif Steil and Sacha A.F.T. van Hijum. „Defining the structure of the general stress regulon of *Bacillus subtilis* using targeted microarray analysis and Random Forest Classification". *Microbiology* 158, 696-707.

Pierre Nicolas, Ulrike Mäder, Etienne Dervyn, Tatiana Rochat, Aurélie Leduc, Nathalie Pigeonneau, Elena Bidnenko, Elodie Marchadier, Mark Hoebeke, Stéphane Aymerich, Dörte Becher, Paola Bisicchia, Eric Botella, Olivier Delumeau, Geoff Doherty, Emma L. Denham, Kevin M. Devine, Mark Fogg, Vincent Fromion, Anne Goelzer, Annette Hansen, Elisabeth Härtig, Colin R. Harwood, Georg Homuth, Hanne Jarmer, Matthieu Jules, Edda Klipp, Ludovic Le Chat, François Lecointe, Peter Lewis, Wolfram Liebermeister, Anika March, Ruben A.T. Mars, **Priyanka Nannapaneni**, David Noone, Susanne Pohl, Bernd Rinn, Frank Rügheimer, Praveen K. Sappa, Franck Samson, Marc Schaffer, Benno Schwikowski, Leif Steil, Jörg Stülke, Thomas Wiegert, Anthony J Wilkinson, Jan Maarten van Dijl, Michael Hecker, Uwe Völker, Philippe Bessières, and Philippe Noirot. „The condition-dependent whole-transcriptome reveals high-level regulatory architecture in bacteria". *Science* 335, 1103-1106.

Poster presentations:

Poster presentation titled ''Transcriptome analysis of *Bacillus subtilis* and definition of the SigB regulon using a random forest class prediction algorithm'' during VAAM 2009 at Bochum, Germany.

ACKNOWLEDGEMENTS

APPENDIX

Appendix 2.1.5 Scripts to train and test the Random Forest models

Training:

```
featurematrix <- read.csv("featurematrix.txt",header=T,row.names=1,sep='\t')
featurevector <- read.csv("featurevector.txt",header=FALSE,sep='\t')
featurematrix$test <-  as.factor(as.character(t(featurevector)))
library(randomForest)
randf <- randomForest(test ~., data=featurematrix,keep.forest=T, ntree=10000 , importance=TRUE,
proximity=TRUE, replace=FALSE, sampsize=ceiling(0.632*nrow(featurematrix)), keep.inbag=FALSE)
print(randf)
save(randf, file="generatedRandomForest.rfo")
save(randf, file="randomForest.rfo")
randf$predicted
varUsed(randf)
write.table(randf$predicted, file="Prediction.txt", sep="\t", dec=".", quote=FALSE,col.names=F)
colrs <- c('red','blue')
if((dim(featurematrix)-1) > 30) {
num = 30;
}else{
num = dim(featurematrix)-1;
}
num
bitmap(file="rForest-VarImp.jpg",height=100,width=30); varImpPlot(randf,main='Variable importance
plot',n.var=(dim(featurematrix)-1),cex=0.7,sort=F);dev.off()
bitmap(file="rForst-MDS.jpg",height=60,width=60,pointsize=18);mds <-
MDSplot(randf,featurematrix$test,palette=colrs,cex=0.7,k=num);dev.off()
clustmeth <- 'complete'
cl<-hclust(as.dist(1-randf$proximity),method=clustmeth)
titl <- 'Hierarchical clustering using Random Forest sample proximities'subtitl=paste("agglomeration
method='",clustmeth,"'",sep='')
bitmap(file="rForest-Clustering.jpg",height=30,width=100,pointsize=30);plot(cl,xlab='',main=titl,
sub=subtitl, labels=paste(as.character(rownames(featurematrix)),"_",featurematrix$test));dev.off()
bitmap(file="rForest-Outliers.jpg",height=10,width=30,pointsize=14);barplot(abs(outlier(randf)),
col=c('red','green')[as.numeric(featurematrix$test)],names.arg=as.character(rownames(featurematrix)))
;dev.off()
bitmap(file="rForest-Errors.jpg",height=10,width=30,pointsize=14);barplot(t(abs(featurevector-
as.numeric(as.matrix(randf$predicted)))),col=c('red','green')[as.numeric(featurematrix$test)],names.arg
=as.character(rownames(featurematrix)));dev.off()
quit()
```

Testing:

```
feature <- read.table("test_matrix.txt",header=T,row.names=1,sep='\t')
library(party)
library(randomForest)
load("randomForest.rfo")
prediction = as.matrix(predict(randf,feature))
prob = as.matrix(predict(randf,feature,type="prob"))
rownames(prediction) = rownames(feature)
prediction
prob
write.table(prediction, file="prediction_testmatrix.txt", sep="\t", dec=".", quote=FALSE, col.names=F)
write.table(prob, file="probabilities.txt", sep="\t", dec=".", quote=FALSE,col.names=F)
```

```
featurematrix <- read.csv("featurematrix.txt",header=T,row.names=1,sep='\t')
featurevector <- read.csv("featurevector.txt",header=FALSE,sep='\t')
featurematrix$test <-  as.factor(as.character(t(featurevector)))
library(randomForest)
randf <- randomForest(test ~., data=featurematrix,keep.forest=T, importance=TRUE,
proximity=TRUE,localImp=TRUE)
write.table(file="localImp.txt",sep="\t",quote=F,row.names=F,randf$localImportance)
```

Appendix 3.1.1: All genes on the array (350 genes)

Gene[a]	BSU number	Description[b]	Comment
abnA	BSU28810	endo-1,5-alpha-L-arabinosidase	found in (3)
ahpC	BSU40090	alkyl hydroperoxide reductase (small subunit)	heat shock dependent gene in (1)
aldY	BSU38830	aldehyde dehydrogenase	found in (1, 3)
aroK	BSU03150	shikimate kinase	putative genes in (2)
bglS	BSU39070	endo-beta-1,3-1,4 glucanase	found in (3)
bioW	BSU30240	6-carboxyhexanoate-CoA ligase	genes from reported operons
bioA	BSU30230	adenosylmethionine-8-amino-7-oxononanoate aminotransferase	found in (3)
bioB	BSU30200	biotin synthetase	genes from reported operons
bmrU	BSU24000	multidrug resistance protein cotranscribed with bmr	found in (1, 2, 3)
bmr	BSU24010	multidrug-efflux transporter	found in (2)
bmrR	BSU24020	transcriptional activator of the bmrUR	found in (2, 3)
csbX	BSU27760	sigma-B-transcribed gene	found in (1, 2, 3)
bofC	BSU27750	forespore regulator of the sigma-K checkpoint	found in (1, 2)
citB	BSU18000	aconitate hydratase	control
citG	BSU33040	fumarate hydratase	control
citZ	BSU29140	citrate synthase II (major)	found in (3)
mdh	BSU29120	malate dehydrogenase	control
ctsR	BSU00830	transcriptional repressor of class III stress genes	found in (2, 3)
mcsA	BSU00840	modulator of CtsR repression	found in (1, 2, 3)
mcsB	BSU00850	modulator of CtsR repression	found in (1, 2, 3)
clpC	BSU00860	class III stress response-related ATPase	found in (2, 3)
radA	BSU00870	DNA repair protein homolog	found in (2, 3)
yacK	BSU00880	unknown; similar to unknown proteins	found in (2)
clpE	BSU13700	ATP-dependent Clp protease-like (class III stress gene)	control
clpP	BSU34540	ATP-dependent Clp protease proteolytic subunit (class III heat-shock protein)	found in (2)
csbA	BSU35180	putative membrane protein	found in (1)
csbB	BSU08600	stress response protein	found in (1)
csbC	BSU39810	putative sugar transporter	found in (1, 2, 3)
csbD	BSU36670	sigma-B-controlled gene	found in (1, 2, 3)
cspB	BSU09100	major cold-shock protein	control
cspC	BSU05120	cold-shock protein	control
ctc	BSU00520	general stress protein	found in (1, 2, 3)
dctB	BSU04440	possible C4-dicarboxylate binding protein	found in (3)
ydbD	BSU04430	unknown; similar to manganese-containing catalase	found in (1, 2, 3)
des	BSU19180	membrane phospholipid desaturase	control
dhbA	BSU32000	2,3-dihydro-2,3-dihydroxybenzoate dehydrogenase	control
dnaK	BSU25470	class I heat-shock protein (molecular chaperone)	control
dps	BSU30650	stress- and starvation-induced gene controlled by sigma-B	found in (1, 2, 3)

Gene[a]	BSU number	Description[b]	Comment
eno	BSU33900	enolase	control
feuA	BSU01630	iron-uptake system (binding protein)	control
fhuD	BSU33320	ferrichrome ABC transporter (ferrichrome-binding protein)	control
gabD	BSU03910	succinate-semialdehyde dehydrogenase	found in (1, 2, 3)
gabP	BSU06310	gamma-aminobutyrate permease	found in (3)
gerE	BSU28410	transcriptional regulator required for the expression of late spore coat genes	control
gerPA	BSU10720	probable spore germination protein	found in (3)
ggaA	BSU35690	galactosamine-containing minor teichoic acid biosynthesis	found in (3)
gpr	BSU25540	spore protease (degradation of SASPs)	control
groEL	BSU06030	class I heat-shock protein (chaperonin)	control
gsiB	BSU04400	general stress protein	found in (1, 2, 3)
gspA	BSU38430	general stress protein	found in (1, 2, 3)
gtaB	BSU35670	UTP-glucose-1-phosphate uridylyltransferase	found in (2)
guaD	BSU13170	guanine deaminase	found in (2, 3)
hisZ	BSU34930	histidyl-tRNA synthetase	genes from reported operons
hisD	BSU34910	histidinol dehydrogenase	found in (3)
hisI	BSU34860	phosphoribosyl-AMP cyclohydrolase / phosphoribosyl-ATP pyrophosphohydrolase	genes from reported operons
hom	BSU32260	homoserine dehydrogenase	control
ilvC	BSU28290	ketol-acid reductoisomerase	control
katA	BSU08820	vegetative catalase 1	control
katE	BSU39050	catalase 2	found in (1, 2, 3)
katX	BSU38630	major catalase in spores	found in (2, 3)
ldh	BSU03050	L-lactate dehydrogenase	found in (1)
lctP	BSU03060	L-lactate permease	genes from reported operons
leuC	BSU28260	3-isopropylmalate dehydratase (large subunit)	control
lexA	BSU17850	transcriptional repressor of the SOS regulon	found in (3)
licR	BSU38600	transcriptional activator of the lichenan	found in (3)
lipA	BSU32330	probable lipoic acid synthetase	found in (3)
lonB	BSU28210	Lon-like ATP-dependent protease	control
malS	BSU29880	malate dehydrogenase	found in (3)
manP	BSU12010	putative PTS mannose-specific enzyme IIBCA component	found in (3)
miaA	BSU17330	tRNA isopentenylpyrophosphate transferase	found in (3)
mmgB	BSU24160	3-hydroxybutyryl-CoA dehydrogenase	genes from reported operons
mmgD	BSU24140	citrate synthase III	found in (3)
mmgE	BSU24130	function unknown	genes from reported operons
mrgA	BSU32990	metalloregulation DNA-binding stress protein	control
mtn	BSU27270	methylthioadenosine nucleosidase	putative genes in (2)
mtrA	BSU22780	GTP cyclohydrolase I	putative genes in (2)
nadC	BSU27860	nicotinate-nucleotide pyrophosphorylase	found in (3)

Gene[a]	BSU number	Description[b]	Comment
nadE	BSU03130	NH3-dependent NAD+ synthetase	found in (2, 3)
nap	BSU05440	carboxylesterase NA	found in (3)
nhaX	BSU09690	putative regulatory gene for nhaC	found in (1, 2, 3)
opuAA	BSU02980	glycine betaine ABC transporter (ATP-binding protein)	control
opuBA	BSU33730	choline ABC transporter (ATP-binding protein)	genes from reported operons
opuBB	BSU33720	choline ABC transporter (membrane protein)	found in (3)
opuBC	BSU33710	choline ABC transporter (choline-binding protein)	found in (3)
opuBD	BSU33700	choline ABC transporter (membrane protein)	genes from reported operons
opuCA	BSU33830	glycine betaine/carnitine/choline ABC transporter (ATP-binding protein)	control
opuD	BSU30070	glycine betaine transporter	found in (3)
opuE	BSU06660	proline transporter	found in (3)
pdhA	BSU14580	pyruvate dehydrogenase (E1 alpha subunit)	control
pfkA	BSU29190	6-phosphofructokinase	control
proH	BSU18480	pyrroline-5-carboxylate reductase	control
ptb	BSU24090	probable phosphate butyryltransferase	found in (3)
purK	BSU06430	phosphoribosylaminoimidazole carboxylase II	putative genes in (2)
rbfA	BSU16650	ribosome-binding factor A	found in (3)
rph	BSU28370	ribonuclease PH	found in (3)
rplJ	BSU01040	ribosomal protein L10 (BL5)	control
rpsF	BSU40910	ribosomal protein S6 (BS9)	control
rsbU	BSU04700	indirect positive regulator of sigma-B activity (serine phosphatase)	found in (1)
rsbV	BSU04710	positive regulator of sigma-B activity (anti-anti-sigma factor)	found in (1, 2)
rsbW	BSU04720	negative regulator of sigma-B activity (switch protein/serine kinase, anti-sigma factor)	found in (1, 2, 3)
sigB	BSU04730	RNA polymerase general stress sigma factor	found in (1, 2, 3)
rsbX	BSU04740	indirect negative regulator of sigma-B activity (serine phosphatase)	found in (1, 2, 3)
rsfA	BSU37620	probable regulator of transcription of sigma-F-dependent genes	found in (3)
sat	BSU15590	probable sulfate adenylyltransferase	control
sdhC	BSU28450	succinate dehydrogenase (cytochrome b558 subunit)	found in (3)
sdhA	BSU28440	succinate dehydrogenase (flavoprotein subunit)	genes from reported operons
sodA	BSU25020	superoxide dismutase	found in (2)
spoIID	BSU36750	required for complete dissolution of the asymmetric septum	control
spoIIID	BSU36420	transcriptional regulator of sigma-E- and sigma-K-dependent genes	control
spoIIQ	BSU36550	required for completion of engulfment	control
spoIIR	BSU36970	required for processing of pro-sigma-E	control
spoVC	BSU00530	probable peptidyl-tRNA hydrolase	found in (2)
spsA	BSU37910	spore coat polysaccharide synthesis	control

Gene[a]	BSU number	Description[b]	Comment
sspE	BSU08660	small acid-soluble spore protein (major gamma-type SASP)	control
trxA	BSU28500	thioredoxin	found in (1, 2, 3)
tufA	BSU01130	elongation factor Tu	control
xkdS	BSU12720	PBSX prophage	found in (3)
xpf	BSU12560	RNA polymerase PBSX sigma factor-like	found in (3)
yaaH	BSU00160	unknown; similar to cortical fragment-lytic enzyme	found in (1, 2)
yaaI	BSU00170	unknown; similar to isochorismatase	found in (1, 2)
yabK	BSU00540	unknown	putative genes in (2)
yabS	BSU00650	unknown; similar to unknown proteins	additional genes of interest
yabT	BSU00660	unknown; similar to serine/threonine-protein kinase	found in (3)
yacL	BSU00890	unknown; similar to unknown proteins	found in (1, 2, 3)
yacM	BSU00900	unknown; similar to unknown proteins	found in (3)
yazC	BSU00950	unknown; similar to unknown proteins	putative genes in (2)
yacO	BSU00960	unknown; similar to tRNA/rRNA methyltransferase	putative genes in (2)
ybyB	BSU02110	unknown; similar to unknown proteins	found in (1, 2, 3)
ycbO	BSU02580	unknown; similar to unknown proteins	found in (3)
ycbP	BSU02590	unknown; similar to unknown proteins from B. subtilis	found in (1, 2, 3)
ycdB	BSU02790	unknown; similar to unknown proteins	found in (3)
ycdF	BSU02830	unknown; similar to glucose 1-dehydrogenase	found in (1, 2, 3)
ycdG	BSU02840	unknown; similar to oligo-1,6-glucosidase	found in (1, 2, 3)
yceC	BSU02890	unknown; similar to tellurium resistance protein	found in (2)
yceD	BSU02900	unknown; similar to tellurium resistance protein	found in (2)
yceE	BSU02910	unknown; similar to tellurium resistance protein	found in (2)
yceF	BSU02920	unknown; similar to tellurium resistance protein	found in (2)
yceG	BSU02930	unknown	found in (2)
yceH	BSU02940	unknown; similar to toxic anion resistance protein	found in (2)
ycsD	BSU04030	unknown; similar to hydroxymyristoyl-(acyl carrier protein) dehydratase	found in (3)
ycsE	BSU04040	unknown; similar to unknown proteins from B. subtilis	putative genes in (2)
ydaB	BSU04170	unknown; similar to unknown proteins	found in (1)
ydaC	BSU04180	unknown; similar to unknown proteins	found in (1)
ydaD	BSU04190	unknown; similar to alcohol dehydrogenase	found in (1, 2, 3)
ydaE	BSU04200	unknown	found in (1, 2, 3)
ydaF	BSU04210	unknown; similar to acetyltransferase	found in (1, 3)
ydaG	BSU04220	unknown; similar to general stress protein	found in (1, 2, 3)
ydaJ	BSU04270	unknown	found in (1, 3)
ydaK	BSU04280	unknown	found in (1)
ydaL	BSU04290	unknown	found in (1)
ydaM	BSU04300	unknown; similar to cellulose synthase	found in (1)
ydaN	BSU04310	unknown	found in (1)
ydaP	BSU04340	unknown; similar to pyruvate oxidase	found in (1, 2, 3)
ydaT	BSU04380	unknown	found in (1, 2, 3)

Gene[a]	BSU number	Description[b]	Comment
ydaS	BSU04370	unknown	found in (1, 2, 3)
ydbB	BSU04410	unknown; similar to unknown proteins	additional genes of interest
ydbC	BSU04420	unknown	found in (3)
ydbP	BSU04550	unknown; similar to thioredoxin	found in (4)
ydcF	BSU04750	unknown	putative genes in (2)
yddS	BSU05090	unknown; similar to metabolite transport protein	putative genes in (2)
ydeC	BSU05150	unknown; similar to transcriptional regulator (AraC/XylS family)	found in (3)
ydfC	BSU05360	unknown; similar to unknown proteins	found in (3)
ydfO	BSU05490	unknown; similar to unknown proteins	found in (3)
ydgC	BSU05580	unknown; similar to unknown proteins	found in (1)
ydhK	BSU05790	unknown	found in (1)
ydjB	BSU06120	unknown	found in (3)
ydjJ	BSU06220	unknown	found in (3)
ydjL	BSU06240	unknown; similar to L-iditol 2-dehydrogenase	found in (3)
yeaC	BSU06330	unknown; similar to methanol dehydrogenase regulation	putative genes in (2)
yebA	BSU06350	unknown; similar to unknown proteins	found in (3)
yerD	BSU06590	unknown; similar to glutamate synthase (ferredoxin)	found in (1, 3)
yesO	BSU06970	unknown; similar to sugar-binding protein	additional genes of interest
yesP	BSU06980	unknown; similar to lactose permease	found in (3)
yesQ	BSU06990	unknown; similar to lactose permease	additional genes of interest
yesR	BSU07000	unknown; similar to unknown proteins	found in (3)
yfhF	BSU08510	unknown; similar to cell-division inhibitor	found in (1, 2, 3)
yfhE	BSU08500	unknown	found in (2, 3)
yfhD	BSU08490	unknown	found in (1, 2)
yfhK	BSU08570	unknown; similar to cell-division inhibitor	found in (1, 2, 3)
yfhL	BSU08580	unknown	found in (1, 2)
yfhM	BSU08590	unknown; similar to epoxide hydrolase	found in (1, 2)
yfhO	BSU08610	unknown	found in (1)
yfjB	BSU08160	unknown	putative genes in (2)
yfkC	BSU07940	unknown; similar to unknown proteins	putative genes in (2)
yfkE	BSU07920	unknown; similar to H+/Ca2+ exchanger	found in (1, 3)
yfkD	BSU07930	unknown; similar to unknown proteins	found in (1, 3)
yfkJ	BSU07880	unknown; similar to protein-tyrosine phosphatase	found in (1, 2, 3)
yfkI	BSU07890	unknown	found in (2, 3)
yfkH	BSU07900	unknown; similar to unknown proteins	found in (1, 2, 3)
yfkM	BSU07850	unknown; similar to unknown proteins	found in (1, 2, 3)
yfkS	BSU07770	unknown	found in (3)
yfkT	BSU07760	unknown; similar to spore germination protein	found in (3)
yflA	BSU07750	unknown; similar to aminoacid carrier protein	found in (1, 2, 3)
yflS	BSU07570	unknown; similar to 2-oxoglutarate/malate translocator	found in (1, 3)

Gene[a]	BSU number	Description[b]	Comment
yflT	BSU07550	unknown	found in (1, 2, 3)
yfmA	BSU07540	unknown	found in (3)
yfmG	BSU07480	unknown	putative genes in (2)
yfmK	BSU07440	unknown	putative genes in (2)
ygaE	BSU08700	unknown; similar to unknown proteins	found in (3)
ygxA	BSU08750	unknown; similar to unknown proteins from B. subtilis	putative genes in (2)
ygxB	BSU09390	unknown	found in (1, 2, 3)
yhaS	BSU09870	unknown	found in (1)
yhaT	BSU09860	unknown; similar to unknown proteins	found in (1)
yhaU	BSU09850	unknown; similar to Na+/H+ antiporter	found in (1)
yhcM	BSU09140	unknown	found in (1, 2)
yhdE	BSU09380	unknown; similar to unknown proteins	putative genes in (2)
yhdF	BSU09450	unknown; similar to glucose 1-dehydrogenase	found in (1, 2)
yhdG	BSU09460	unknown; similar to amino acid transporter	found in (1)
yhdN	BSU09530	unknown; similar to aldo/keto reductase	found in (1, 2)
yhjJ	BSU10530	unknown; similar to unknown proteins	found in (3)
yhxD	BSU10430	unknown; similar to alcohol dehydrogenase	found in (1, 2, 3)
yisP	BSU10810	unknown; similar to phytoene synthase	found in (3)
yjbC	BSU11490	unknown; similar to unknown proteins	found in (1, 2, 3)
yjbD	BSU11500	unknown; similar to unknown proteins	found in (1, 2, 3)
yjcE	BSU11830	unknown	found in (1, 2, 3)
yjdJ	BSU12070	unknown	found in (3)
yjgA	BSU12140	unknown	found in (1)
yjgB	BSU12150	unknown	found in (1, 2, 3)
yjgC	BSU12160	unknown; similar to formate dehydrogenase	found in (1, 2, 3)
yjgD	BSU12170	unknown; similar to unknown proteins from B. subtilis	found in (1, 2, 3)
yjlB	BSU12270	unknown; similar to unknown proteins	found in (3)
ykgA	BSU13020	unknown; similar to unknown proteins	found in (1, 2, 3)
ykgB	BSU13010	unknown; similar to unknown proteins	found in (1, 3)
ykoC	BSU13210	unknown; similar to unknown proteins	found in (3)
ykpA	BSU14430	unknown; similar to ABC transporter (ATP-binding protein)	found in (3)
ykrT	BSU13560	unknown; similar to unknown proteins	putative genes in (2)
ykrS	BSU13550	unknown; similar to initiation factor eIF-2B (alpha subunit)	putative genes in (2)
ykuT	BSU14210	unknown; similar to unknown proteins	found in (2, 3)
ykyB	BSU14020	unknown	putative genes in (2)
ykzA	BSU13160	unknown; similar to organic hydroperoxide resistance protein	found in (1, 2, 3)
ykzC	BSU14680	unknown	putative genes in (2)
ynfC	BSU18110	unknown	found in (3)
yoaN	BSU18670	unknown; similar to unknown proteins	found in (3)
yocB	BSU19150	unknown	found in (1, 2, 3)
yocK	BSU19240	unknown; similar to general stress protein	found in (1, 2, 3)

Gene[a]	BSU number	Description[b]	Comment
yocL	BSU19250	unknown	additional genes of interest
yodB	BSU19540	unknown; similar to unknown proteins	found in (3)
yorA	BSU20450	unknown; similar to unknown proteins from B. subtilis	found in (3)
yotK	BSU19850	unknown	found in (4)
yoxA	BSU18360	unknown; similar to unknown proteins from B. subtilis	found in (3)
yoxC	BSU18510	unknown	found in (1, 2, 3)
yoxB	BSU18520	unknown	found in (1, 3)
yoaA	BSU18530	unknown; similar to ribosomal-protein-alanine N-acetyltransferase	found in (1)
ypuB	BSU23340	unknown	found in (4)
ypuC	BSU23330	unknown	found in (1)
ypuD	BSU23300	unknown	found in (1, 2)
ypzE	BSU23060	unknown	found in (3)
yqaS	BSU26200	unknown; similar to phage-related terminase small subunit	additional genes of interest
yqaT	BSU26190	unknown; similar to phage-related terminase large subunit	additional genes of interest
yqbA	BSU26180	unknown; similar to phage-related protein	found in (3)
yqbB	BSU26170	unknown	additional genes of interest
yqgE	BSU25010	unknown; similar to unknown proteins	found in (3)
yqgZ	BSU24770	unknown; similar to unknown proteins	found in (1, 2, 3)
yqhA	BSU24760	unknown; similar to positive regulator of sigma-B activity	found in (1, 3)
yqhB	BSU24750	unknown; similar to hemolysin	found in (1, 2, 3)
yqhQ	BSU24490	unknown; similar to unknown proteins	found in (4)
yqhP	BSU24500	unknown; similar to unknown proteins	found in (1)
yqiY	BSU23970	unknown; similar to amino acid ABC transporter (permease)	found in (3)
yqiZ	BSU23960	unknown; similar to amino acid ABC transporter (ATP-binding protein)	genes from reported operons
yqjF	BSU23900	unknown; similar to unknown proteins	found in (3)
yqjL	BSU23830	unknown; similar to unknown proteins from B. subtilis	found in (2)
yqxH	BSU25910	unknown; similar to holin	found in (3)
yqxL	BSU24740	unknown; similar to unknown proteins	found in (1, 2, 3)
yraA	BSU27020	unknown; similar to general stress protein	found in (2)
yraE	BSU26980	unknown; similar to spore coat protein	found in (3)
yrhD	BSU27230	unknown; similar to unknown proteins from B. subtilis	found in (3)
yrhE	BSU27220	unknown; similar to formate dehydrogenase	additional genes of interest
yrkE	BSU26540	unknown; similar to unknown proteins	found in (3)
yrvC	BSU27640	unknown; similar to unknown proteins	found in (3)
yrvD	BSU27630	unknown; similar to unknown proteins	found in (4)
ysdB	BSU28830	unknown; similar to unknown proteins	found in (1)
yshC	BSU28590	unknown; similar to DNA polymerase beta	found in (3)

Gene[a]	BSU number	Description[b]	Comment
ysnF	BSU28340	unknown; similar to unknown proteins	found in (2)
ysxD	BSU28180	unknown	found in (3)
ytaB	BSU30930	unknown	found in (1, 2, 3)
ytiA	BSU30700	unknown; similar to ribosomal protein	found in (2, 3)
ytkL	BSU29410	unknown; similar to unknown proteins	found in (1, 3)
ytxG	BSU29780	unknown; similar to general stress protein	found in (1, 2, 3)
ytxH	BSU29770	unknown; similar to general stress protein	found in (1, 2, 3)
ytxJ	BSU29760	unknown; similar to general stress protein	found in (1, 2, 3)
ytzB	BSU29870	unknown; similar to unknown proteins	putative genes in (2)
ytzE	BSU30020	unknown; similar to transcriptional regulator (DeoR family)	found in (3)
yugU	BSU31280	unknown; similar to unknown proteins	found in (2)
yuiD	BSU32060	unknown; similar to unknown proteins	found in (3)
yumB	BSU32100	unknown; similar to NADH dehydrogenase	putative genes in (2)
yunG	BSU32400	unknown	found in (1)
yurH	BSU32530	unknown; similar to N-carbamyl-L-amino acid amidohydrolase	genes from reported operons
yurG	BSU32520	unknown; similar to aspartate aminotransferase	found in (3)
yusD	BSU32760	unknown	putative genes in (2)
yusO	BSU32870	unknown; similar to transcriptional regulator (MarR family)	genes from reported operons
yusP	BSU32880	unknown; similar to multidrug-efflux transporter	found in (3)
yusS	BSU32910	unknown; similar to 3-oxoacyl- acyl-carrier protein reductase	putative genes in (2)
yutC	BSU32320	unknown; similar to unknown proteins	found in (3)
yutK	BSU32180	unknown; similar to Na+/nucleoside cotransporter	putative genes in (2)
yuzA	BSU31380	unknown; similar to unknown proteins	found in (1, 2, 3)
yvaA	BSU33530	unknown; similar to unknown proteins	found in (1, 2, 3)
yvaB	BSU33540	unknown; similar to NAD(P)H dehydrogenase (quinone)	found in (3)
yvaC	BSU33550	unknown; similar to unknown proteins	found in (3)
yvaG	BSU33590	unknown; similar to 3-oxoacyl- acyl-carrier protein reductase	found in (1)
yvaK	BSU33620	unknown; similar to carboxylesterase	found in (1, 2, 3)
rnr	BSU33610	ribonuclease R	found in (1, 2, 3)
yvaM	BSU33640	unknown; similar to hydrolase	putative genes in (2)
yvaN	BSU33660	unknown; similar to immunity repressor protein	found in (3)
yvbG	BSU33850	unknown; similar to unknown proteins	found in (1, 3)
yvfB	BSU34265	unknown	genes from reported operons
yvfD	BSU34240	unknown; similar to serine O-acetyltransferase	found in (3)
yvfF	BSU34220	unknown; similar to unknown proteins from B. subtilis	genes from reported operons
yvgN	BSU33400	unknown; similar to dehydrogenase	found in (1, 2)
yvgO	BSU33410	unknown; similar to unknown proteins	found in (1, 2)
yvgT	BSU33460	unknown; similar to unknown proteins	found in (3)
yvgW	BSU33490	unknown; similar to heavy metal-transporting ATPase	found in (1)

Gene[a]	BSU number	Description[b]	Comment
yvgZ	BSU33520	unknown; similar to unknown proteins	found in (1)
yvqJ	BSU33140	unknown; similar to macrolide-efflux protein	found in (3)
yvrE	BSU33200	unknown; similar to senescence marker protein-30	found in (1, 2, 3)
yvyD	BSU35310	unknown; similar to sigma-54 modulating factor of gram-negative bacteria	found in (1, 2, 3)
ywaF	BSU38440	unknown	found in (3)
ywdD	BSU38000	unknown	found in (1)
ywdJ	BSU37940	unknown; similar to unknown proteins	putative genes in (2)
ywdL	BSU37920	unknown; similar to unknown proteins	putative genes in (2)
ywfH	BSU37680	unknown; similar to 3-oxoacyl- acyl-carrier protein reductase	found in (3)
ywhH	BSU37480	unknown; similar to unknown proteins	found in (3)
ywiE	BSU37240	unknown; similar to cardiolipin synthetase	found in (1, 2, 3)
ywjA	BSU37230	unknown; similar to ABC transporter (ATP-binding protein)	found in (3)
ywjB	BSU37220	unknown; similar to unknown proteins	found in (3)
ywjC	BSU37210	unknown	found in (2, 3)
ywlB	BSU36960	unknown; similar to unknown proteins	putative genes in (2)
ywmE	BSU36720	unknown	found in (2, 3)
ywmF	BSU36680	unknown; similar to unknown proteins	found in (3)
ywsB	BSU35970	unknown; similar to unknown proteins from B. subtilis	found in (2, 3)
ywsC	BSU35900	unknown; similar to capsular polyglutamate biosynthesis	found in (1, 3)
ywtA	BSU35890	unknown; similar to capsular polyglutamate biosynthesis	found in (3)
ywtG	BSU35830	unknown; similar to metabolite transport protein	found in (1, 2, 3)
ywzA	BSU38180	unknown; similar to unknown proteins from B. subtilis	found in (2, 3)
yxaB	BSU40030	unknown; similar to unknown proteins	found in (2, 3)
yxaC	BSU40022	unknown; similar to unknown proteins	found in (3)
yxbG	BSU39840	unknown; similar to glucose 1-dehydrogenase	found in (1, 2, 3)
yxjG	BSU38960	unknown; similar to unknown proteins from B. subtilis	found in (3)
yxjJ	BSU38930	unknown	found in (3)
yxjL	BSU38910	unknown; similar to two-component response regulator [YxjM]	genes from reported operons
yxkO	BSU38720	unknown; similar to unknown proteins	found in (1, 2)
yxlJ	BSU38620	unknown; similar to DNA-3-methyladenine glycosidase	found in (1, 2, 3)
yxzF	BSU38610	unknown	found in (1, 2)
yxnA	BSU40000	unknown; similar to glucose 1-dehydrogenase	found in (1, 2, 3)
yxxB	BSU39440	unknown	found in (3)
yyaM	BSU40810	unknown; similar to unknown proteins	found in (3)
yyaN	BSU40800	unknown; similar to transcriptional regulator (MerR family)	additional genes of interest
yybO	BSU40570	unknown; similar to ABC transporter (permease)	found in (3)
yycD	BSU40450	unknown	found in (2, 3)
yycE	BSU40430	unknown	found in (1)

Gene[a]	BSU number	Description[b]	Comment
yycR	BSU40250	unknown; similar to formaldehyde dehydrogenase	found in (3)

[a] Genes are sorted based on their gene name, downstream genes in (putative) operons are indented.
[b] Descriptions are derived from subtiwiki (http://subtiwiki.uni-goettingen.de)

1. **Helmann, J. D., Wu, M. F., Kobel, P. A., Gamo, F. J., Wilson, M., Morshedi, M. M., Navre, M. & Paddon, C. (2001).** Global transcriptional response of *Bacillus subtilis* to heat shock. *J Bacteriol* **183**, 7318-7328.
2. **Petersohn, A., Brigulla, M., Haas, S., Hoheisel, J. D., Völker, U. & Hecker, M. (2001).** Global analysis of the general stress response of *Bacillus subtilis. J Bacteriol* **183**, 5617-5631.
3. **Price, C. W., Fawcett, P., Ceremonie, H., Su, N., Murphy, C. K. & Youngman, P. (2001).** Genome-wide analysis of the general stress response in *Bacillus subtilis. Mol Microbiol* **41**, 757-774.
4. **Petersohn, A., Bernhardt, J., Gerth, U., Höper, D., Koburger, T., Völker, U. & Hecker, M. (1999).** Identification of sigma(B)-dependent genes in *Bacillus subtilis* using a promoter consensus-directed search and oligonucleotide hybridization. *J Bacteriol* **181**, 5718-5724.

Appendix 3.1.2: SigB regulon members (166) predicted from "expression RF model"

Gene[a]	BSU number	Description[b]	"expression RF model"
bioA	BSU30230	lysine-8-amino-7-oxononanoate aminotransferase	Predicted SigB
bmr	BSU24010	multidrug-efflux transporter	Predicted SigB
bmrU	BSU24000	multidrug resistance protein cotranscribed with bmr	Identified in (1, 2, 3)
bofC	BSU27750	forespore regulator of the sigma-K checkpoint	Predicted SigB
clpC	BSU00860	class III stress response-related ATPase	Predicted SigB
clpP	BSU34540	ATP-dependent Clp protease proteolytic subunit (class III heat-shock protein)	Predicted SigB
csbA	BSU35180	putative membrane protein	Predicted SigB
csbC	BSU39810	putative sugar transporter	Identified in (1, 2, 3)
csbD	BSU36670	sigma-B-controlled gene	Identified in (1, 2, 3)
csbX	BSU27760	sigma-B-transcribed gene	Identified in (1, 2, 3)
ctc	BSU00520	general stress protein	Identified in (1, 2, 3)
ctsR	BSU00830	transcriptional repressor of class III stress genes	Predicted SigB
dps	BSU30650	stress- and starvation-induced gene controlled by sigma-B	Identified in (1, 2, 3)
gabD	BSU03910	succinate-semialdehyde dehydrogenase	Identified in (1, 2, 3)
ggaA	BSU35690	galactosamine-containing minor teichoic acid biosynthesis	Predicted SigB
gsiB	BSU04400	general stress protein	Identified in (1, 2, 3)
gspA	BSU38430	general stress protein	Identified in (1, 2, 3)
gtaB	BSU35670	UTP-glucose-1-phosphate uridylyltransferase	Predicted SigB
guaD	BSU13170	guanine deaminase	Predicted SigB
katE	BSU39050	catalase 2	Identified in (1, 2, 3)
lctP	BSU03060	L-lactate permease	Predicted SigB
mcsA	BSU00840	modulator of CtsR repression	Identified in (1, 2, 3)
mcsB	BSU00850	modulator of CtsR repression	Identified in (1, 2, 3)
mmgE	BSU24130	function unknown	Predicted SigB
nadE	BSU03130	NH3-dependent NAD+ synthetase	Predicted SigB
nhaX	BSU09690	putative regulatory gene for nhaC	Identified in (1, 2, 3)
opuBD	BSU33700	choline ABC transporter (membrane protein)	Predicted SigB
opuE	BSU06660	proline transporter	Predicted SigB
ptb	BSU24090	probable phosphate butyryltransferase	Predicted SigB
rbfA	BSU16650	ribosome-binding factor A	Predicted SigB
rnr	BSU33610	ribonuclease R	Identified in (1, 2, 3)
rph	BSU28370	ribonuclease PH	Predicted SigB
rsbV	BSU04710	positive regulator of sigma-B activity (anti-anti-sigma factor)	Predicted SigB
rsbW	BSU04720	negative regulator of sigma-B activity (switch protein/serine kinase, anti-sigma factor)	Identified in (1, 2, 3)
rsbX	BSU04740	indirect negative regulator of sigma-B activity (serine phosphatase)	Identified in (1, 2, 3)
sigB	BSU04730	RNA polymerase general stress sigma factor	Identified in (1, 2, 3)
sodA	BSU25020	superoxide dismutase	Predicted SigB
spoVC	BSU00530	probable peptidyl-tRNA hydrolase	Predicted SigB
trxA	BSU28500	thioredoxin	Identified in (1, 2, 3)

Gene[a]	BSU number	Description[b]	"expression model"	RF
xkdS	BSU12720	PBSX prophage	Predicted SigB	
yaaH	BSU00160	unknown; similar to cortical fragment-lytic enzyme	Predicted SigB	
yaaI	BSU00170	unknown; similar to isochorismatase	Predicted SigB	
yabK	BSU00540	unknown	Predicted SigB	
yacL	BSU00890	unknown; similar to unknown proteins	Identified in (1, 2, 3)	
ybyB	BSU02110	unknown; similar to unknown proteins	Identified in (1, 2, 3)	
ycbO	BSU02580	unknown; similar to unknown proteins	Predicted SigB	
ycbP	BSU02590	unknown; similar to unknown proteins from B. subtilis	Identified in (1, 2, 3)	
ycdF	BSU02830	unknown; similar to glucose 1-dehydrogenase	Identified in (1, 2, 3)	
ycdG	BSU02840	unknown; similar to oligo-1,6-glucosidase	Identified in (1, 2, 3)	
ycsD	BSU04030	unknown; similar to hydroxymyristoyl-(acyl carrier protein) dehydratase	Predicted SigB	
ycsE	BSU04040	unknown; similar to unknown proteins from B. subtilis	Predicted SigB	
ydaB	BSU04170	unknown; similar to unknown proteins	Predicted SigB	
ydaC	BSU04180	unknown; similar to unknown proteins	Predicted SigB	
ydaD	BSU04190	unknown; similar to alcohol dehydrogenase	Identified in (1, 2, 3)	
ydaE	BSU04200	unknown	Identified in (1, 2, 3)	
ydaG	BSU04220	unknown; similar to general stress protein	Identified in (1, 2, 3)	
ydaJ	BSU04270	unknown	Predicted SigB	
ydaK	BSU04280	unknown	Predicted SigB	
ydaM	BSU04300	unknown; similar to cellulose synthase	Predicted SigB	
ydaN	BSU04310	unknown	Predicted SigB	
ydaP	BSU04340	unknown; similar to pyruvate oxidase	Identified in (1, 2, 3)	
ydaS	BSU04370	unknown	Identified in (1, 2, 3)	
ydaT	BSU04380	unknown	Identified in (1, 2, 3)	
ydbD	BSU04430	unknown; similar to manganese-containing catalase	Identified in (1, 2, 3)	
ydbP	BSU04550	unknown; similar to thioredoxin	Predicted SigB	
ydfC	BSU05360	unknown; similar to unknown proteins	Predicted SigB	
ydgC	BSU05580	unknown; similar to unknown proteins	Predicted SigB	
ydhK	BSU05790	unknown	Predicted SigB	
ydjJ	BSU06220	unknown	Predicted SigB	
yeaC	BSU06330	unknown; similar to methanol dehydrogenase regulation	Predicted SigB	
yerD	BSU06590	unknown; similar to glutamate synthase (ferredoxin)	Predicted SigB	
yesP	BSU06980	unknown; similar to lactose permease	Predicted SigB	
yfhD	BSU08490	unknown	Predicted SigB	
yfhE	BSU08500	unknown	Predicted SigB	
yfhF	BSU08510	unknown; similar to cell-division inhibitor	Identified in (1, 2, 3)	
yfhK	BSU08570	unknown; similar to cell-division inhibitor	Identified in (1, 2, 3)	
yfhL	BSU08580	unknown	Predicted SigB	
yfhM	BSU08590	unknown; similar to epoxide hydrolase	Predicted SigB	
yfhO	BSU08610	unknown	Predicted SigB	
yfkD	BSU07930	unknown; similar to unknown proteins	Predicted SigB	
yfkE	BSU07920	unknown; similar to H+/Ca2+ exchanger	Predicted SigB	
yfkH	BSU07900	unknown; similar to unknown proteins	Identified in (1, 2, 3)	

Gene[a]	BSU number	Description[b]	"expression model"	RF
yfkI	BSU07890	unknown	Predicted SigB	
yfkJ	BSU07880	unknown; similar to protein-tyrosine phosphatase	Identified in (1, 2, 3)	
yfkM	BSU07850	unknown; similar to unknown proteins	Identified in (1, 2, 3)	
yfkS	BSU07770	unknown	Predicted SigB	
yfkT	BSU07760	unknown; similar to spore germination protein	Predicted SigB	
yflA	BSU07750	unknown; similar to aminoacid carrier protein	Identified in (1, 2, 3)	
yflS	BSU07570	unknown; similar to 2-oxoglutarate/malate translocator	Predicted SigB	
yflT	BSU07550	unknown	Identified in (1, 2, 3)	
ygxB	BSU09390	unknown	Identified in (1, 2, 3)	
yhcM	BSU09140	unknown	Predicted SigB	
yhdF	BSU09450	unknown; similar to glucose 1-dehydrogenase	Predicted SigB	
yhdN	BSU09530	unknown; similar to aldo/keto reductase	Predicted SigB	
yhxD	BSU10430	unknown; similar to alcohol dehydrogenase	Identified in (1, 2, 3)	
yjbC	BSU11490	unknown; similar to unknown proteins	Identified in (1, 2, 3)	
yjbD	BSU11500	unknown; similar to unknown proteins	Identified in (1, 2, 3)	
yjcE	BSU11830	unknown	Identified in (1, 2, 3)	
yjgA	BSU12140	unknown	Predicted SigB	
yjgB	BSU12150	unknown	Identified in (1, 2, 3)	
yjgC	BSU12160	unknown; similar to formate dehydrogenase	Identified in (1, 2, 3)	
yjgD	BSU12170	unknown; similar to unknown proteins from B. subtilis	Identified in (1, 2, 3)	
yjlB	BSU12270	unknown; similar to unknown proteins	Predicted SigB	
ykgA	BSU13020	unknown; similar to unknown proteins	Identified in (1, 2, 3)	
ykgB	BSU13010	unknown; similar to unknown proteins	Predicted SigB	
ykyB	BSU14020	unknown	Predicted SigB	
ykzA	BSU13160	unknown; similar to organic hydroperoxide resistance protein	Identified in (1, 2, 3)	
ykzC	BSU14680	unknown	Predicted SigB	
ynfC	BSU18110	unknown	Predicted SigB	
yoaA	BSU18530	unknown; similar to ribosomal-protein-alanine N-acetyltransferase	Predicted SigB	
yocB	BSU19150	unknown	Identified in (1, 2, 3)	
yocK	BSU19240	unknown; similar to general stress protein	Identified in (1, 2, 3)	
yorA	BSU20450	unknown; similar to unknown proteins from B. subtilis	Predicted SigB	
yoxB	BSU18520	unknown	Predicted SigB	
yoxC	BSU18510	unknown	Identified in (1, 2, 3)	
ypuC	BSU23330	unknown	Predicted SigB	
yqbA	BSU26180	unknown; similar to phage-related protein	Predicted SigB	
yqgZ	BSU24770	unknown; similar to unknown proteins	Identified in (1, 2, 3)	
yqhA	BSU24760	unknown; similar to positive regulator of sigma-B activity	Predicted SigB	
yqhB	BSU24750	unknown; similar to hemolysin	Identified in (1, 2, 3)	
yqhP	BSU24500	unknown; similar to unknown proteins	Predicted SigB	
yqhQ	BSU24490	unknown; similar to unknown proteins	Predicted SigB	
yqiZ	BSU23960	unknown; similar to amino acid ABC transporter (ATP-binding protein)	Predicted SigB	

Gene[a]	BSU number	Description[b]	"expression model"	RF
yqjL	BSU23830	unknown; similar to unknown proteins from B. subtilis	Predicted SigB	
yqxL	BSU24740	unknown; similar to unknown proteins	Identified in (1, 2, 3)	
yrkE	BSU26540	unknown; similar to unknown proteins	Predicted SigB	
yrvD	BSU27630	unknown; similar to unknown proteins	Predicted SigB	
yshC	BSU28590	unknown; similar to DNA polymerase beta	Predicted SigB	
ytaB	BSU30930	unknown	Identified in (1, 2, 3)	
ytiA	BSU30700	unknown; similar to ribosomal protein	Predicted SigB	
ytkL	BSU29410	unknown; similar to unknown proteins	Predicted SigB	
ytxG	BSU29780	unknown; similar to general stress protein	Identified in (1, 2, 3)	
ytxH	BSU29770	unknown; similar to general stress protein	Identified in (1, 2, 3)	
ytxJ	BSU29760	unknown; similar to general stress protein	Identified in (1, 2, 3)	
ytzB	BSU29870	unknown; similar to unknown proteins	Predicted SigB	
yutK	BSU32180	unknown; similar to Na+/nucleoside cotransporter	Predicted SigB	
yuzA	BSU31380	unknown; similar to unknown proteins	Identified in (1, 2, 3)	
yvaA	BSU33530	unknown; similar to unknown proteins	Identified in (1, 2, 3)	
yvaB	BSU33540	unknown; similar to NAD(P)H dehydrogenase (quinone)	Predicted SigB	
yvaG	BSU33590	unknown; similar to 3-oxoacyl- acyl-carrier protein reductase	Predicted SigB	
yvaK	BSU33620	unknown; similar to carboxylesterase	Identified in (1, 2, 3)	
yvbG	BSU33850	unknown; similar to unknown proteins	Predicted SigB	
yvfF	BSU34220	unknown; similar to unknown proteins from B. subtilis	Predicted SigB	
yvgO	BSU33410	unknown; similar to unknown proteins	Predicted SigB	
yvgZ	BSU33520	unknown; similar to unknown proteins	Predicted SigB	
yvrE	BSU33200	unknown; similar to senescence marker protein-30	Identified in (1, 2, 3)	
yvyD	BSU35310	unknown; similar to sigma-54 modulating factor of gram-negative bacteria	Identified in (1, 2, 3)	
ywiE	BSU37240	unknown; similar to cardiolipin synthetase	Identified in (1, 2, 3)	
ywjA	BSU37230	unknown; similar to ABC transporter (ATP-binding protein)	Predicted SigB	
ywjB	BSU37220	unknown; similar to unknown proteins	Predicted SigB	
ywjC	BSU37210	unknown	Predicted SigB	
ywsB	BSU35970	unknown; similar to unknown proteins from B. subtilis	Predicted SigB	
ywtA	BSU35890	unknown; similar to capsular polyglutamate biosynthesis	Predicted SigB	
ywtG	BSU35830	unknown; similar to metabolite transport protein	Identified in (1, 2, 3)	
ywzA	BSU38180	unknown; similar to unknown proteins from B. subtilis	Predicted SigB	
yxaB	BSU40030	unknown; similar to unknown proteins	Predicted SigB	
yxbG	BSU39840	unknown; similar to glucose 1-dehydrogenase	Identified in (1, 2, 3)	
yxjJ	BSU38930	unknown	Predicted SigB	
yxjL	BSU38910	unknown; similar to two-component response regulator [YxjM]	Predicted SigB	
yxkO	BSU38720	unknown; similar to unknown proteins	Predicted SigB	
yxlJ	BSU38620	unknown; similar to DNA-3-methyladenine glycosidase	Identified in (1, 2, 3)	
yxnA	BSU40000	unknown; similar to glucose 1-dehydrogenase	Identified in (1, 2, 3)	
yxxB	BSU39440	unknown	Predicted SigB	
yxzF	BSU38610	unknown	Predicted SigB	

Gene[a]	BSU number	Description[b]	"expression model"	RF
yybO	BSU40570	unknown; similar to ABC transporter (permease)	Predicted SigB	
yycD	BSU40450	unknown	Predicted SigB	

[a] Genes are sorted based on gene name
[b] Descriptions are derived from subtiwiki (http://subtiwiki.uni-goettingen.de)

1. **Helmann, J. D., Wu, M. F., Kobel, P. A., Gamo, F. J., Wilson, M., Morshedi, M. M., Navre, M. & Paddon, C. (2001).** Global transcriptional response of *Bacillus subtilis* to heat shock. *J Bacteriol* **183**, 7318-7328.
2. **Petersohn, A., Brigulla, M., Haas, S., Hoheisel, J. D., Völker, U. & Hecker, M. (2001).** Global analysis of the general stress response of *Bacillus subtilis*. *J Bacteriol* **183**, 5617-5631.
3. **Price, C. W., Fawcett, P., Ceremonie, H., Su, N., Murphy, C. K. & Youngman, P. (2001).** Genome-wide analysis of the general stress response in *Bacillus subtilis*. *Mol Microbiol* **41**, 757-774.

Appendix 3.1.3: SigB regulon members (37) with secondary regulators predicted from „kinetic RF model"

Gene[a]	BSU number	Description[b]	Known secondary Regulator[b]	Reference[b]
bioB	BSU30200	biotin synthase	BirA regulon	(Bower *et al.*, 1996; Perkins *et al.*, 1996)
bmrR	BSU24020	general stress protein, transcriptional activator of the bmrU-bmr-bmrR operon	SigB, bmrR, Mta	(Baranova *et al.*, 1999; Kumaraswami *et al.*, 2010; Petersohn *et al.*, 1999a; Petersohn *et al.*, 2001)
*nap**	BSU05440	carboxylesterase NA		
opuBA	BSU33730	choline ABC transporter (ATP-binding protein)		
opuBB	BSU33720	choline ABC transporter (membrane protein)		
purK	BSU06430	phosphoribosylaminoimidazole carboxylase (ATP-dependent)	G-box, PurR regulon	(Ebbole & Zalkin, 1987; Weng *et al.*, 1995)
radA	BSU00870	suppressor of recU and recB mutations, participates in the stabilization and or processing of holliday junction intermediates	CtsR regulon, SigB regulon	(Derre *et al.*, 1999; Petersohn *et al.*, 2001)
yacK	BSU00880	DNA integrity scanning protein, delays sporulation in the case of chromosome damage, has diadenylate cyclase activity	SigM, CtsR	(Jervis *et al.*, 2007; Krüger *et al.*, 1996)
yacM	BSU00900	2-C-methyl-D-erythritol 4-phosphate cytidylyltransferase	SigB, SigM	(Höper *et al.*, 2005; Jervis *et al.*, 2007)
yceC	BSU02890	similar to tellurium resistance protein	SigB, SigW, SigM	(Cao *et al.*, 2002; Petersohn *et al.*, 2001)
yceD	BSU02900	general stress protein, similar to tellurium resistance protein	SigB, SigW, SigM	(Cao *et al.*, 2002; Petersohn *et al.*, 2001)
yceE	BSU02910	general stress protein, similar to tellurium resistance protein	SigB, SigW, SigM	(Cao *et al.*, 2002; Petersohn *et al.*, 2001)
yceF	BSU02920	general stress protein, similar to tellurium resistance protein	SigB, SigW, SigM	(Cao *et al.*, 2002; Petersohn *et al.*, 2001)
yceG	BSU02930	general stress protein, survival of ethanol stress	SigB, SigW, SigM	(Cao *et al.*, 2002; Petersohn *et al.*, 2001)
yceH	BSU02940	similar to toxic anion resistance protein	SigB, SigW, SigM	(Cao *et al.*, 2002; Petersohn *et al.*, 2001)
ydaF	BSU04210	similar to acetyltransferase	SigB, AbrB	(Banse *et al.*, 2008)
ydbB	BSU04410	unknown		
ydjL	BSU06240	acetoine/ butanediol dehydrogenase	AbrB regulon	
yfmG	BSU07480	unknown	AbrB regulon	
yhaS	BSU09870	modulator of YhaU activity		
yhaT	BSU09860	K+/H+ antiporter for K+ efflux		
yhaU	BSU09850	K+/H+ antiporter for K+ efflux		
yisP	BSU10810	unknown		

Gene[a]	BSU number	Description[b]	Known secondary Regulator[b]	Reference[b]
*ykrT**	BSU13560	5-methylthioribose kinase	SigA, S-box	(Grundy & Henkin, 1998; Sekowska & Danchin, 2002)
*yocL**	BSU19250	unknown	SigE regulon	
ypuD	BSU23300	unknown	LexA regulon, SigB regulon, SigM regulon	(Au *et al.*, 2005; Eiamphungporn & Helmann, 2008; Petersohn *et al.*, 2001)
*yqgE**	BSU25010	unknown		
yqiY	BSU23970	high affinity arginine ABC transporter (permease)		
yqxH	BSU25910	similar to holin		
yraA	BSU27020	general stress protein, degradation of damaged thiol-containing proteins	AdhR regulon, SigB regulon, SigM regulon	(Jervis *et al.*, 2007; Nguyen *et al.*, 2009)
ysnF	BSU28340	general stress protein, survival of ethanol stress		
*ytzE**	BSU30020	similar to transcriptional regulator (DeoR family)		
yvaC	BSU33550	unknown		
*ywaF**	BSU38440	unknown		
*ywhH**	BSU37480	unknown	ComA regulon	(Comella & Grossman, 2005)
ywmE	BSU36720	general stress protein, survival of ethanol stress		
yxjG	BSU38960	unknown	S-box	(Grundy & Henkin, 1998)

[a] Genes are sorted based on gene name

[b] Descriptions are derived from subtiwiki (http://subtiwiki.uni-goettingen.de)

* These genes were not considered to be SigB-dependent, as we could not identify putative SigB-dependent promoter upstream of these genes.

Appendix 3.1.4: Promoters of the SigB regulon members either known or predicted in the study

Gene[a]	Potential promoter sequence	Source[b]
bioA	**TGTTAA**CCTTTGAATATAATT**GGTTAA**	Motif finder prediction
bioB	Potential *bioW-bioA-bioB* operon	
bmrU	**GTTTAC**TGCTTACAGAAAAA**GGGGAT**	DBTBS
bmr	Potential *bmrU-bmr-bmrR* operon	
bmrR	Potential *bmrU-bmr-bmrR* operon	
csbX	**GATTA**CAATTCAGCAAGCTT**GGGTACA**	DBTBS
bofC	Potential *csbX-bofC* operon	
ctsR	**GTTTTG**TGGACCGGGAAAA**TGGAAAT**	DBTBS
mcsA	Potential *ctsR-mscA-mcsB-clpC-radA-yacK* operon	
mcsB	Potential *ctsR-mscA-mcsB-clpC-radA-yacK* operon	
clpC	Potential *ctsR-mscA-mcsB-clpC-radA-yacK* operon	
radA	Potential *ctsR-mscA-mcsB-clpC-radA-yacK* operon	
yacK	Potential *ctsR-mscA-mcsB-clpC-radA-yacK* operon	
clpP	**GTTTGA**ACCCCTGTATTTTTG**GGGAAA**	DBTBS
csbA	**GATTGA**TTTTGGCTGAAAAG**GGGTAT**	DBTBS
csbC	**GTTTCA**AATGAGATAGGAAAT**GGGTAC**	DBTBS
csbD	**GTTTA**TTGCCTCTCAGATC**GGGAAG**	DBTBS
ctc	**GTTTAA**ATCCTTATCGTTAT**GGGTAT**	DBTBS
dps	**GTTTTA**GCGTACGATATTAA**GGGTAT**	DBTBS
gabD	**GATTAC**TTTTGCTGACAGC**GGGAAT**	DBTBS
ggaA	**ATTTTC**AGCAAAAAACGCC**AGGAAA**	Motif finder prediction
gsiB	**GTTTAA**AAGAATTGTGAGC**GGGAAT**	DBTBS
gspA	**GTTTAT**TTTTTTGAAAAA**GGGTAT**	DBTBS
gtaB	**GTGTAA**AAACATATTGAAAA**GGGTAA**	DBTBS
katE	**GTTTAT**ATGAAGAACGCCAC**GGGTAA**	DBTBS
mmgE	**GTATAC**AGTCAATCGTCCGGA**AGATAT**	Motif finder prediction
nadE	**GATTCA**TTTTCATTGATTTA**GGGAAA**	DBTBS
opuBA	**TTATAA**ACTTTATTCTATAAT**GGGAAG**	Motif finder prediction
opuBB	Potential *opuBA-opuBB-opuBC-opuBD* operon	
opuBD	Potential *opuBA-opuBB-opuBC-opuBD* operon	
opuE	**GTTTCA**TCGCTTGAATGGCC**GGGAAT**	DBTBS
purK	**ACATTA**TCCATGTCCGTTGTTA**AGATAA**	Motif finder prediction
rbfA	**AGTTTC**TGTCACTTGCTT**AGGTAT**	Motif finder prediction
rph	**GGTTTA**TGTACATACAATCGGA**GGTAAA**	Motif finder prediction
rsbV	**GTTTAA**CGTCTGTCAGACGA**GGGTAT**	DBTBS
rsbW	Potential *rsbV-rsbW-sigB-rsbX* operon	
rsbX	Potential *rsbV-rsbW-sigB-rsbX* operon	
sigB	Potential *rsbV-rsbW-sigB-rsbX* operon	
sodA	**GTTTAA**TGGGGCAGATTATC**GGTTAAA**	found in (2)
trxA	**GTTTTA**AAACAGCTCCGGCA**GGGCAT**	DBTBS
xkdS	**TCGTTT**TTTATTTTAGATA**AGGTAT**	Motif finder prediction
yaaI	**GTTTT**TTCATTGCCTAAAAA**GGCTA**	found in (1, 2)
yacL	**GGTTAA**AAACCTTATGAATAC**GGGTAT**	DBTBS
yacM	Potential *yacL-yacM* operon	DBTBS
ybyB	**GTTTA**GCAATTTCCAAAAC**GGGAA**	found in (1, 2)
ycbO	**ACTTTA**AAATATTAAATG**AGAAAA**	Motif finder prediction
ycbP	**GTTTAA**CTTTTTACATTTGA**GGAAT**	found in (1, 2)
ycdF	**GTTTT**CAACTCGGAAAAACA**GGGTAT**	found in (2)
ycdG	Potential *ycdF-ycdG* operon	
yceC	**GTATAT**ATTAGTAATTTAC**GGCTTA**TTTT	found in (2)
yceD	Potential *yceC-yceD-yceF-yceG-yceH* operon	
yceE	Potential *yceC-yceD-yceF-yceG-yceH* operon	
yceF	Potential *yceC-yceD-yceF-yceG-yceH* operon	
yceG	Potential *yceC-yceD-yceF-yceG-yceH* operon	

Gene[a]	Potential promoter sequence	Source[b]
yceH	Potential *yceC-yceD-yceF-yceG-yceH* operon	
ycsD	**TCGTTT**CACTTTAGGCTG**AGATAT**	Motif finder prediction
ycsE	**TCGTAA**TCCCGTTCGCTTTGAG**AGTTAT**	Motif finder prediction
ydaB	**GTTTTA**CAAGTTATCTTTTTT**GGGTTA**	found in (1)
ydaC	**GTTTCA**CATGGAACGCTGAGA**GGAAA**	found in (1)
ydaD	**GTTTAT**CCAAAGAGTGTGTG**AGGTAC**	DBTBS
ydaE	Potential *ydaD-ydaE-ydaF* operon	
ydaF	Potential *ydaD-ydaE-ydaF* operon	
ydaG	**GTTTAA**ATCTTCCCCGGATG**TGGAAA**	
ydaJ	**GTTTT**CTTAATGTTCAAAAA**GGGAAA**	found in (1)
ydaK	**GTTTTA**GCGGACTGGATCAAGC**AGGAAT**	Motif finder prediction
ydaM	**TTATTT**GTCATCATGTTTGTC**AGTTAT**	Motif finder prediction
ydaN	**TGTTAT**CTACTCTTTATTCGT**AGAAAT**	Motif finder prediction
ydaP	**GTTTTA**AAGCCTTTCTCCTG**TGGTAT**	DBTBS
ydaT	**GTTTTT**ATTTTTCACCTGC**GGGTAC**	DBTBS
ydaS	Potential *ydaT-ydaS* operon	
ydbB	**TGATAA**AGAATTCTACCA**AGAAAT**	Motif finder prediction
ydbD	**TCGTTT**ATCTTTCTATCGA**TCGGAAA**	DBTBS
ydbP	**GATTAC**AGCCAGAGCTGTCTGCA**GGGAAA**	DBTBS
ydfC	**TTTTTC**ATCTTTTTTTCTTCT**GGAAAG**	Motif finder prediction
ydgC	**GTGAGT**AACATTACTCGT**GGGTAT**	found in (1)
ydhK	**GTTTGG**CTTTGCAAACAAAG**GGGAAT**	DBTBS
ydjJ	**GTTTCA**GACTCAGGATAAAA**GGAAAT**	Motif finder prediction
ydjL	**TTTTTT**ATTTCAATATCCTTAC**AGTAAT**	Motif finder prediction
yeaC	**TTATAC**AGATTGCTTTTTTGT**GGTAAA**	Motif finder prediction
yerD	**GTTTG**GAAAGTGTTCTACTGT**GGAAA**	found in (1)
yesP	**ACGTTC**CGGAAAAAGGCCAATG**AGATAT**	Motif finder prediction
yfhF	**GTTTT**CTTTTATTACAATGA**GGTAA**	found in (1, 2)
yfhE	Potential *yfhF-yfhE-yfhD* operon	
yfhD	Potential *yfhF-yfhE-yfhD* operon	
yfhK	**GTTTCA**CCAGCCTGTCAATCA**GGGAAT**	DBTBS
yfhL	Potential *yfhK-yfhL-yfhM* operon	
yfhM	Potential *yfhK-yfhL-yfhM* operon	
yfkE	**GTTTT**CCAAAAGCAGGCAACCT**GAAAA**	found in (1)
yfkJ	**GTTT**CTTTTTAGAGAAATAG**GGGCAA**	DBTBS
yfkI	Potential *yfkJ-yfkI-yfkH* operon	DBTBS
yfkH	Potential *yfkJ-yfkI-yfkH* operon	DBTBS
yfkM	**GTTTAT**TTTAGTGTGGCTG**GGGTAG**	found in (2)
yflA	**GTTTAT**GTTTTTCCATCTAT**GGGAAA**	DBTBS
yflS	**GTTAA**GGAGTAGAATGGAAAA**GGGGAT**	found in (1)
yflT	**GTTTCA**GGTACAGACGATC**GGGTAT**	found in (2)
yfmG	**TCTTTA**TACTAATGAAAAAT**GGTAAA**	Motif finder prediction
ygxB	**GTATAA**ATAATTTCAGCC**GGGCAG**	found in (1, 2)
yhaS	**GTTTTA**TATAGTGAAAAAGAA**GGGATA**	found in (1)
yhaT	Potential *yhaS-yhaT* operon	
yhaU	**GATTAT**CAAACACAACCA**GGAAAA**	Motif finder prediction
yhcM	**GGTTAA**TTTGTCTAACGAGG**GGGAAA**	found in (1, 2)
yhdF	**GTTTAT**TCATTTCTGTCGTG**TGGTAA**	DBTBS
yhdN	**GTTTAA**CATTTTTTCCAGAG**GGGAAA**	DBTBS
yhxD	**GTTTT**CTGCTTATGCTCAG**GGGTA**	found in (1, 2)
yisP	**GCATTT**TCTTCATATACAATTA**GGAAAA**	Motif finder prediction
yjbC	**GTTTAA**ACAAGAAGAAATG**GGGTAT**	DBTBS
yjbD	Potential *yjbC-yjbD* operon	
yjcE	**GTTTT**ACAAGAAACAC**GGGTAT**	found in (1, 2)
yjgA	**TCATAA**AGACATGACAATTTC**AGAAAT**	Motif finder prediction
yjgB	**GTTTCG**TTGCAAAAACACG**GGGAAA**	DBTBS
yjgC	**GTTTT**ATTGAGTTGTTGTAA**GGGAAC**	found in (1, 2)
yjgD	Potential *yjgC-yjgD* operon	

Gene[a]	Potential promoter sequence	Source[b]
ykgA	**GTTTAA**TGATTTTCATGATGA**GGGAAT**	DBTBS
ykgB	Potential *ykgA-ykgB* operon	DBTBS
ykzA	**GTTTAA**AAAGATCAGAAA**GGGAAA**	DBTBS
ynfC	**GTTTTT**ATTCACACCATAAACC**GGGTAT**	Motif finder prediction
yocB	**GTTTGA**TCGTTTTTAAGAGA**GGAAAA**	found in (1, 2)
yocK	**GTTTGA**CAGAAGGCAAAAC**GGGAAA**	DBTBS
yorA	**TGATTT**AACAATCTCAATTCAT**GGGTAT**	Motif finder prediction
yoxC	**GATTAA**AAAAACGGATACA**GGGTAA**	DBTBS
yoxB	Potential *yoxC-yoxB-yoaA* operon	
yoaA	Potential *yoxC-yoxB-yoaA* operon	
ypuC	**GATTAT**ACAAAAAGTGGATT**GGGAAT**	DBTBS
ypuD	**GTTTTT**TATTCATGAAAAAAA**GGAAT**	found in (1, 2)
yqbA	**TGGTAC**GCTCTCGATGATTATG**AGGTAA**	Motif finder prediction
yqgZ	**TGGTT**TAAATGAAAAATGAT**CCGGGTA**	DBTBS
yqhA	**GTTTTT**CACCTGTCCAGAAACTG**GGCTAG**	DBTBS
yqhB	**GTTTT**ATGAGCATTTTCAGGT**GGTAT**	found in (1, 2)
yqhQ	**GAGTAA**AATTTGAAAATAAC**GGGTAT**	DBTBS
yqhP	Potential *yqhQ-yqhP* operon	
yqiY	**TTTTTT**GTGCATAAATTTCCAA**GGATAT**	Motif finder prediction
yqiZ	Potential *yqiY-yqiZ* operon	
yqjL	**GTTCAA**TACGAAAGAGGCT**GGTAAT**	found in (2)
yqxH	**ACGTTT**GCCACTAGCACAAAA**GGAAAA**	Motif finder prediction
yqxL	**GTTTAG**TGACGCGGTTATT**GGGCAA**	DBTBS
yraA	**ATTTAC**AACACTTATGATGA**AGAAAG**	Motif finder prediction
yrkE	**GTGTAT**TGACTATGATGGAACA**AGGAAA**	Motif finder prediction
yrvD	**GTTTAA**AACCAGTCTTGATT**GGGAAA**	DBTBS
yshC	**ACATAT**GTGAACGTCTGGG**GGTTAT**	Motif finder prediction
ysnF	**GTTTAA**TTCAAAGAACAGC**GGGAAT**	found in (2)
ytaB	**GTTTGA**TATTTATAAGATAAA**GGGTAA**	found in (1, 2)
ytiA	**GTTTAC**GATGTGAAACAGA**GGGAAG**	found in (2)
ytkL	**GTTTTT**CAGCTTTTTAAAAA**GGGAAA**	DBTBS
ytxG	**GTTTAT**GATTGAAGAAAC**GGGTAA**	DBTBS
ytxH	Potential *ytxG-ytxH-ytxJ* operon	
ytxJ	Potential *ytxG-ytxH-ytxJ* operon	
yutK	**ATATTA**TTGTTATCGTATAAA**AGATAT**	Motif finder prediction
yuzA	**GTTTTAA**TAATTCATGGA**GGAGG**	found in (1, 2)
yvaA	**GTTTTA**CCATTTGATCAGGA**GGGTAT**	found in (1, 2)
yvaB	**GCTTTT**GAAAAGTAAGCTGTTT**GGGAAA**	Motif finder prediction
yvaC	**GTGTGG**TGGCTGTTTAAT**GGGTAT**	Motif finder prediction
yvaG	**GATTT**CTGTCAATAAATAAGA**GGAAT**	found in (1)
yvaK	**GTTTT**TTTCTGATTAAACTGT**GGAAAA**	found in (2)
rnr	Potential *yvaK-rnr-smpB-ssrA* operon	
yvbG	**GTTTA**CCGGGAAATCGCCTCC**GGGTAA**	found in (1)
yvfD	**TTATAA**AATTTAAAATAATAA**GGGAAG**	Motif finder prediction
yvfF	Potential *yvfD-yvfF* operon	
yvgO	**GATTA**CAAATACATTGAGCA**GGGTAT**	found in (1, 2)
yvgZ	**GTTTG**GACAATCAGTATAAT**GGGAAT**	found in (1)
yvrE	**GTTTGG**ACACCTCTTTGCC**GGGAAT**	DBTBS
yvyD	**GTTTCA**GCAGGAATTGTAAA**GGGTAA**	DBTBS
ywiE	**GTTTAT**CGATTAGAAAAAAGA**GGTAAT**	found in (1, 2)
ywjA	**GTTTTC**AGGTGTTCTGTA**AGAAAA**	Motif finder prediction
ywjC	**AGGTTT**ACGACTTGTCAGCTT**TGGGAAC**	DBTBS
ywmE	**TTGGTT**TAAAAACAGTTTGGGC**GGGAAT**	DBTBS
ywsB	**GTTTAG**GAACCTGCGATAAC**GTGAAT**	found in (2)
ywtA	**GTTTG**GCTTAGTCGATTA**GGGAA**	found in (1)
ywtG	**AGGTTT**AATGGCCGGAAAAAGA**GGCTAA**	DBTBS
ywzA	**GTTTAT**CTTATACAAAAAAGA**GGGAAT**	found in (2)
yxaB	**GCATAGC**CATCCTTCTTTTTT**GGGTAG**	DBTBS

Gene[a]	Potential promoter sequence	Source[b]
yxbG	**GTTTAT**CACTGCACATAGC**GGGAA**	found in (1, 2)
yxjG	**GCTTTT**TTATATTTGAAT**GGAAAG**	Motif finder prediction
yxjJ	**GGGTAT**AAACCAAACAAGCA**AGAAAA**	Motif finder prediction
yxjL	**ACTTAT**GGATTTGGCCTGAC**AGGAAT**	Motif finder prediction
yxkO	**GTTTGA**AAAAGAAAAGGGAC**AGGAAA**	DBTBS
yxlJ	**GTTTT**TTTTGATCTGCTTC**GGGAA**	found in (1, 2)
yxzF	**GTTTTT**TTTGATCTGCTTC**GGGAAT**	Motif finder prediction
yxnA	**GTAAGA**CCCTTCCGGATG**GGGTAA**	found in (1, 2)
yycD	**GTTTCG**GACAGTAACAAGGC**GGGAAA**	DBTBS

[a] Genes are sorted based on gene name, downstream genes in (putative) operons are indented

[b] Promoters were either listed in DBTBS (http://dbtbs.hgc.jp), found by in (1, 2) or predicted by Motif finder search in the current study

1. **Helmann, J. D., Wu, M. F., Kobel, P. A., Gamo, F. J., Wilson, M., Morshedi, M. M., Navre, M. & Paddon, C. (2001).** Global transcriptional response of *Bacillus subtilis* to heat shock. *J Bacteriol* **183**, 7318-7328.
2. **Petersohn, A., Brigulla, M., Haas, S., Hoheisel, J. D., Völker, U. & Hecker, M. (2001).** Global analysis of the general stress response of *Bacillus subtilis. J Bacteriol* **183**, 5617-5631.

Appendix 3.1.5: List of genes in the hierarchical cluster with indicated cluster C1-C4

Gene[a]	BSU number	Expression rank[b]	Promoter rank[b]	Cluster
gsiB	BSU04400	-2.082	-1.703	C1
yhdN	BSU09530	-1.753	-1.627	C1
ykzA	BSU13160	-1.646	-1.627	C1
gspA	BSU38430	-1.850	-1.417	C1
katE	BSU39050	-1.688	-1.322	C1
ykgA	BSU13020	-1.004	-1.703	C1
rsbV	BSU04710	-1.175	-1.570	C1
yflA	BSU07750	-1.073	-1.436	C1
yvaA	BSU33530	-1.125	-1.398	C1
ctc	BSU00520	-0.346	-1.570	C1
ynfC	BSU18110	-0.221	-1.226	C1
ycbP	BSU02590	-0.929	-1.207	C1
ytxG	BSU29780	-0.616	-1.322	C1
ywzA	BSU38180	-0.562	-1.150	C1
opuE	BSU06660	-0.725	-1.055	C1
yjgC	BSU12160	-1.696	-0.692	C1
yfhK	BSU08570	-1.665	-1.055	C1
yflT	BSU07550	-1.797	-0.959	C1
ywiE	BSU37240	-1.802	-0.902	C1
yocK	BSU19240	-0.902	-0.940	C1
yfkM	BSU07850	-0.988	-0.826	C1
yoxC	BSU18510	-1.301	-0.635	C1
ydaG	BSU04220	-1.036	-0.654	C1
ydaP	BSU04340	-0.929	-0.539	C1
yhdF	BSU09450	-0.655	-0.387	C1
ytiA	BSU30700	-0.491	-0.578	C1
ydaT	BSU04380	-0.236	-0.425	C1
bmrU	BSU24000	-0.803	-0.139	C1
ydhK	BSU05790	-0.298	-0.196	C1
ydaC	BSU04180	-0.604	-0.215	C1
yvrE	BSU33200	-0.492	-0.196	C1
ysnF	BSU28340	1.136	-1.703	C2
dps	BSU30650	1.461	-1.398	C2
yjcE	BSU11830	0.362	-1.398	C2
yjbC	BSU11490	0.202	-1.570	C2
yrvD	BSU27630	0.307	-1.627	C2
ytkL	BSU29410	0.392	-1.245	C2
ydaJ	BSU04270	0.444	-1.169	C2
yvbG	BSU33850	0.559	-1.188	C2
yqhB	BSU24750	1.199	-0.539	C2
ctsR	BSU00830	0.912	-0.330	C2
yhaS	BSU09870	1.051	-0.368	C2
yvyD	BSU35310	1.038	-0.883	C2
clpP	BSU34540	0.602	-0.940	C2
csbD	BSU36670	0.796	-0.978	C2
ypuD	BSU23300	0.535	-1.074	C2
ycdF	BSU02830	0.604	-1.131	C2
ydaK	BSU04280	0.634	-1.093	C2
gtaB	BSU35670	-0.269	-0.768	C2
ytaB	BSU30930	-0.183	-0.807	C2
yacL	BSU00890	0.587	-0.749	C2
yhcM	BSU09140	0.107	-0.787	C2
yvaK	BSU33620	0.351	-0.864	C2
yvaB	BSU33540	0.305	-0.559	C2
trxA	BSU28500	0.209	-0.501	C2

Gene[a]	BSU number	Expression rank[b]	Promoter rank[b]	Cluster
yfhF	BSU08510	0.224	-0.482	C2
yqgZ	BSU24770	-1.955	1.693	C3
ywjC	BSU37210	-1.176	1.445	C3
ywtG	BSU35830	-0.678	1.559	C3
yshC	BSU28590	-0.659	1.349	C3
yxnA	BSU40000	-0.904	1.235	C3
ydgC	BSU05580	-0.915	1.273	C3
yhxD	BSU10430	-0.763	0.624	C3
yerD	BSU06590	-0.606	0.662	C3
ydaN	BSU04310	-0.712	0.987	C3
ydfC	BSU05360	-0.619	0.872	C3
yqhA	BSU24760	-0.631	0.930	C3
yqhQ	BSU24490	-2.011	0.319	C3
yqxL	BSU24740	-1.207	0.224	C3
ywjA	BSU37230	-0.882	0.262	C3
yfkJ	BSU07880	-0.841	0.376	C3
yfkE	BSU07920	-0.890	0.472	C3
yjgB	BSU12150	-1.261	-0.005	C3
ydaD	BSU04190	-1.275	0.109	C3
yocB	BSU19150	-1.099	0.071	C3
yycD	BSU40450	-1.000	-0.005	C3
yvgO	BSU33410	-0.934	0.052	C3
yhaU	BSU09850	1.213	0.205	C4
csbC	BSU39810	1.406	0.281	C4
yvaG	BSU33590	0.505	0.434	C4
rph	BSU28370	0.881	0.357	C4
ggaA	BSU35690	0.963	0.453	C4
yxkO	BSU38720	0.012	-0.082	C4
yrkE	BSU26540	-0.171	0.090	C4
nadE	BSU03130	-0.037	0.243	C4
csbA	BSU35180	0.045	0.300	C4
yvgZ	BSU33520	0.426	-0.196	C4
ydaB	BSU04170	0.885	-0.024	C4
ydbP	BSU04550	0.686	0.033	C4
gabD	BSU03910	0.659	-0.062	C4
ydjJ	BSU06220	0.749	-0.043	C4
opuBA	BSU33730	1.838	0.491	C4
ydjL	BSU06240	1.683	0.586	C4
yfmG	BSU07480	1.967	0.853	C4
yisP	BSU10810	1.701	0.739	C4
ygxB	BSU09390	1.655	0.777	C4
ybyB	BSU02110	1.293	1.025	C4
yvaC	BSU33550	0.992	0.605	C4
yxbG	BSU39840	0.716	0.682	C4
ywsB	BSU35970	0.781	0.701	C4
sodA	BSU25020	0.870	0.911	C4
yxlJ	BSU38620	1.166	0.758	C4
yxjG	BSU38960	0.993	0.720	C4
yxjL	BSU38910	0.939	0.796	C4
ydbB	BSU04410	1.094	1.311	C4
ywmE	BSU36720	1.151	1.426	C4
yqxH	BSU25910	1.288	1.368	C4
yceC	BSU02890	1.499	1.578	C4
csbX	BSU27760	1.529	1.464	C4
ydbD	BSU04430	1.674	1.521	C4
xkdS	BSU12720	-0.031	1.197	C4
yxaB	BSU40030	-0.086	1.063	C4
yaaI	BSU00170	0.059	1.044	C4

Gene[a]	BSU number	Expression rank[b]	Promoter rank[b]	Cluster
yeaC	BSU06330	0.341	0.949	C4
yorA	BSU20450	0.462	0.891	C4
yqjL	BSU23830	0.194	1.006	C4
yxjJ	BSU38930	0.360	1.101	C4
ydaM	BSU04300	0.458	1.082	C4
yesP	BSU06980	0.525	1.540	C4
ycsD	BSU04030	0.283	1.407	C4
yuzA	BSU31380	0.120	1.254	C4
yjgA	BSU12140	0.233	1.292	C4
yqbA	BSU26180	0.828	1.330	C4
ycsE	BSU04040	0.730	1.387	C4
yflS	BSU07570	0.939	1.159	C4
yutK	BSU32180	0.724	1.139	C4
ycbO	BSU02580	0.752	1.178	C4

[a] Genes are sorted based on the clusters

[b] The mean of all expression ratio ranks and the corresponding promoter ranks used for the hierarchical clustering were provided as Z-transformed values.

Appendix 3.2.1: The previously un-annotated SigB-dependent RNA features (133 features) with their start and end positions on the genome, the list of conditions in which these features are highly expressed, and their antisense targets (if any).

Feature name[a]	Locus tag[a]	Start position[a]	End position[a]	Classification[a,b]	Antisense name[a]	TU short sig[a]	Highly expressed conditions[a,c]
NA	BSU_misc_RNA_10	698369	698471	-	*NA*	SigA, SigB	MG+10, MG+15, BMM, MG+25, GM+45, MG+5, HPh, M9-exp, MG+t5, GM+90
S322	new_1004835_1004923	1004835	1004923	3'UTR	*NA*	SigB, SigA	Sw, Etha, Salt, G150, Gly, Diami, T0.30H, B60, BC, Oxctl
S323	new_1006655_1006728	1006655	1006728	inter	*NA*	SigB, SigEF, Sig-, SigB, SigA	G150, Etha, Oxctl, H2O2, Gly, Diami, Paraq, Salt, B36, BT
S324	new_1008553_1008622	1008553	1008622	inter	*NA*	SigA, SigB	Etha, H2O2, B36, dia15, Paraq, G150, LoTm, S8, Oxctl, Heat
S329	new_1023253_1023349	1023253	1023349	5'	*NA*	SigA, SigA, SigB	Etha, T2.30H, T-0.40H, T2.0H, T1.30H, S1, GM-0.2, T3.0H, T-1.40H, M9-stat
S339	new_1044247_1044317_c	1044247	1044317	5'	*NA*	SigA, SigB	B60, G135, Etha, LBGexp, G150, Diami, Paraq, G180, H2O2, Gly
S45	new_108559_108673	108559	108673	intra	*NA*	SigWXY, SigA, SigB	Etha, Salt, Heat, dia15, Diami, M9-stat, B60, T-0.40H, T-1.10H, HiTm
S358	new_1090331_1090400	1090331	1090400	intra	*NA*	SigB, SigA	G135, G150, G180, BI, Salt, Etha, Cold, Heat, Oxctl, LBGexp
S365	new_1116857_1117712_c	1116857	1117712	inter	*comK*	SigB	Salt, Etha, C30, Heat, S8, LPhT, Gly, C90, S7, UNK1
S396	new_1189532_1189645	1189532	1189645	inter	*NA*	SigH, SigB, SigA	Etha, S8, Sw, M9-stat, S7, BC, Gly, S6, Salt, C30
S446	new_1289139_1289257	1289139	1289257	inter	*NA*	SigB	Etha, Salt, Heat, S8, M9-stat, S7, LPhT, G135, C30, Sw
S450	new_1298224_1298337	1298224	1298337	inter	*NA*	SigB	Etha, M9-stat, B60, Salt, Gly, LPhT, Heat, Sw, HiTm, B36
S451	new_1298339_1298611	1298339	1298611	5'	*NA*	SigA, SigB	LBGtran, dia5, Lbexp, LBGexp, dia0, HPh, aero, LPh, BMM, MG+25
S464	new_1369127_1369854_c	1369127	1369854	3'PT	*ykfD*	SigB	Etha, S7, Gly, M40t90, C30, H2O2, M0t90, dia15, T1.0H, S8
S468	new_1382628_1383305	1382628	1383305	3'UTR	*guaD*	SigB	Etha, Salt, M9-stat, Sw, B60, LPhT, B36, Heat, Gly, BC

Feature name[a]	Locus tag[a]	Start position[a]	End position[a]	Classification [a,b]	Antisense name[a]	TU short sig[a]	Highly expressed conditions[a,c]
S489	new_1430539_1430623	1430539	1430623	5'	NA	SigB	Etha, Gly, BC, Salt, Sw, B60, LoTm, M9-stat, LPhT, HiTm
S490	new_1430625_1430683	1430625	1430683	5'	NA	SigA, SigB	Etha, Gly, Sw, BC, LoTm, B36, T3.30H, Salt, T5.0H, B60
S497	new_1445542_1445637	1445542	1445637	intra	NA	SigA, SigB, SigK	BT, S8, B36, S7, Etha, S6, G/S, LBGstat, Pyr, Salt
S536	new_1537029_1537112	1537029	1537112	5'	NA	SigK, SigB	Etha, S7, M9-stat, S6, S8, BT, Gly, Salt, B60, Sw
S537	new_1538693_1539646	1538693	1539646	3'PT	yktD	SigA, SigB	H2O2, dia15, Diami, Paraq, G180, G135, G150, dia5, Heat, LBGexp
S555	new_1581224_1582676_c	1581224	1582676	Indep-NT	ftsL	SigB	Salt, Sw, B60, Etha, M9-stat, Gly, C30, LPhT, BC, B36
S572	new_1643910_1644485_c	1643910	1644485	Indep-NT	priA	SigB	Etha, Salt, Sw, Gly, C30, M9-stat, B60, LPhT, HiTm, Heat
S573	new_1655375_1655445	1655375	1655445	intra	NA	SigEF, SigEF, SigB, SigA, SigA	S4, S5, S7, S8, S6, S3, B36, G/S, B60, BT
S9	new_17238_17833_c	17238	17833	Indep-NT	dacA	SigB	Etha, Gly, C30, Salt, S8, S6, HiTm, S7, Sw, C90
S594	new_1738824_1738940	1738824	1738940	5'	NA	SigA, SigA, SigB, SigA	SMM, GM+15, Salt, T-4.40H, GM+45, T-5.40H, MG+45, LPhT, MG+t5
S595	new_1739238_1739323	1739238	1739323	inter	NA	SigA, SigA, SigB, SigA	Cold, LoTm, Oxctl, G180, H2O2, G150, dia15, Paraq, LPh, G135
S596	new_1741529_1741617	1741529	1741617	inter	NA	SigA, SigA, SigB, SigA	G135, G180, G150, H2O2, Oxctl, LBGexp, Paraq, LoTm, LPh, dia15
S680	new_2002172_2002505	2002172	2002505	5'	yoeA	SigB	M9-stat, Etha, Gly, S8, Sw, S7, B60, S6, Salt, C30
S686	new_2021145_2021230	2021145	2021230	3'UTR	yoaB	SigB	Salt, Sw, B60, LoTm, B36, M9-stat, BC, Lbstat, HiTm, Etha
S687	new_2021232_2021971	2021232	2021971	3'PT	yoaB	SigB	LoTm, Salt, Sw, S6, Etha, M9-stat, LPhT, B60, T2.0H, T2.30H
S736	new_2096129_2096321_c	2096129	2096321	inter	NA	SigA, SigB	dia5, Etha, dia15, Diami, S8, S7, C30, G180, G135, C90
S749	new_2119128_2119278_c	2119128	2119278	5'	NA	SigB	S8, S5, S7, S6, Etha, S4, M9-stat, S3, C30, Gly

Feature name[a]	Locus tag[a]	Start position[a]	End position[a]	Classification[a,b]	Antisense name[a]	TU short sig[a]	Highly expressed conditions[a,c]
S809	new_2273515_2273748	2273515	2273748	indep	NA	SigA, SigB	B36, B60, BT, G/S, T4.0H, Pyr, S4, BC, Lbstat, T3.30H
S873	new_2409565_2409634	2409565	2409634	5'	NA	SigB, SigK	S6, S8, S7, S5, BT, B36, C30, Pyr, B60, G/S
S890	new_2448499_2448590_c	2448499	2448590	inter	NA	SigA, SigB, SigEF, SigA	S3, S4, M0t90, G180, S7, S8, Glucon, G135, S5, LoTm
S917	new_2528267_2528403_c	2528267	2528403	intra	NA	SigA, SigB, SigA	H2O2, G180, G135, Cold, G150, Paraq, Oxctl, LBGstat, LPh, HPh
S918	new_2529726_2529888_c	2529726	2529888	inter	NA	SigB, SigA	G180, G150, Oxctl, Etha, G135, GM+120, dia0, HPh, MG+150, LBGexp
S926	new_2560260_2560488_c	2560260	2560488	3'UTR	NA	SigB	Etha, Salt, Sw, Gly, LPhT, B36, Heat, B60, LoTm, M9-stat
S927	new_2562914_2563888	2562914	2563888	3'PT	rsbRD	SigB	Sw, Salt, Etha, LPhT, M9-stat, S8, BC, S6, BT, Gly
S928	new_2563803_2563987_c	2563803	2563987	inter	NA	SigB	Etha, LoTm, Salt, Heat, Gly, M9-stat, B60, S8, S0, Mal
S931	new_2569624_2570653	2569624	2570653	Indep-NT	yqgS	SigB	Etha, Gly, M9-stat, Salt, S8, LPhT, S7, Sw, HiTm, B60
S936	new_2586107_2586175_c	2586107	2586175	5'	NA	SigB, SigA	Etha, Salt, Heat, LPhT, Gly, Sw, S4, LoTm, M9-stat, B36
S937	new_2586704_2586996_c	2586704	2586996	5'	yqgB	SigB, SigA	Etha, Gly, Salt, aero, S0, LoTm, HiTm, LPh, nit, ferm
S957	new_2609894_2610040_c	2609894	2610040	intra	NA	SigA, SigA, SigB	Lbstat, T1.30H, T2.0H, M9-stat, S2, T3.0H, S5, M0t90, T2.30H, T4.0H
S958	new_2611388_2611455_c	2611388	2611455	inter	NA	SigA, SigB	Oxctl, H2O2, G180, G150, GM+120, S1, G135, Diami, LBGtran, C30
S968	new_2647725_2647872_c	2647725	2647872	indep	NA	SigK, SigA, SigB	S8, S7, S6, Etha, LBGstat, T0.30H, Salt, BT, LPhT, C30
S970	new_2648688_2648796_c	2648688	2648796	5'	NA	SigB	Etha, Salt, Gly, S8, C30, S7, LPhT, Sw, Heat, HiTm
S975	new_2664732_2666054	2664732	2666054	Indep-NT	yqxH	SigB, SigEF	M9-stat, LoTm, S4, B60, S3, T1.0H, T0.30H, S5, T1.30H, S6
S1021	new_2773647_2773762	2773647	2773762	inter	NA	SigB	Etha, Salt, LPhT, HiTm, G135, LBGstat, G180, G150, Heat, H2O2
S97	new_280556_281740_c	280556	281740	3'PT	ycbO	SigB	Etha, Sw, Gly, B36, LoTm, B60, GM+150, Salt, C30, ferm

Feature name [a]	Locus tag[a]	Start position[a]	End position[a]	Classification [a,b]	Antisense name[a]	TU short sig[a]	Highly expressed conditions[a,c]
S1045	new_2822771_2822861_c	2822771	2822861	inter	*NA*	SigK, SigA, SigB	Cold, G135, G180, G150, H2O2, Paraq, Oxctl, C30, LBGexp, BMM
S1050	new_2836807_2836908_c	2836807	2836908	inter	*NA*	SigGF, SigB	Etha, Heat, S8, S7, S4, S3, S6, S5, Salt, LoTm
S1063	new_2879205_2879294_c	2879205	2879294	5'	*NA*	SigA, SigB, SigGF	dia15, Salt, dia5, Etha, Paraq, H2O2, Diami, Heat, LoTm, S5
S1064	new_2879296_2879712_c	2879296	2879712	5'	*ysxD*	SigB, SigGF	Etha, S5, Gly, S7, S4, S8, S6, H2O2, dia15, G135
S1078	new_2913380_2913589_c	2913380	2913589	5'	*NA*	SigB, SigA	Etha, Salt, Gly, Sw, M9-stat, LPhT, BC, B60, S5, LoTm
S1112	new_2996680_2996979	2996680	2996979	5'	*nrnA*	SigB, SigK, SigA	Lbstat, S8, S7, S6, S2, Etha, nit, ferm, BT, M0t90
S1134	new_3035475_3036308_c	3035475	3036308	3'PT	*rpsD*	SigB	Etha, S7, Gly, S8, M9-stat, Sw, G135, S6, BC, C30
S1136	new_3036310_3036529_c	3036310	3036529	indep	*rpsD*	SigB	S5, Etha, S8, S6, S7, S4, BT, S3, M9-stat, B36
S1139	new_3045221_3045329_c	3045221	3045329	5'	*NA*	SigA, SigA, SigA, SigB	Oxctl, H2O2, Paraq, MG-0.1, MG-0.2, G150, GM+15, Diami, LBGexp, G180
S1140	new_3045331_3045396_c	3045331	3045396	inter	*NA*	SigA, SigA, SigB	BMM, GM+120, Diami, GM+90, M/G, GM+25, M9-exp, HiTm, G180, Oxctl
S1141	new_3046584_3046684_c	3046584	3046684	5'	*NA*	SigA, SigB, SigH	Lbtran, BMM, dia5, GM+25, GM+90, M9-exp, MG+150, Diami, MG-0.2, SMM
S1153	new_3076730_3076817	3076730	3076817	5'	*NA*	SigB, SigA	Etha, LBGstat, Salt, M9-stat, Lbstat, LBGtran, Sw, C30, Gly, LoTm
S1170	new_3135504_3135638_c	3135504	3135638	5'	*NA*	SigGF, SigK, SigB	BT, B36, S8, S5, S7, Etha, S6, S4, B60, LoTm
S1171	new_3136119_3136205_c	3136119	3136205	inter	*NA*	SigB	Etha, B36, S8, B60, S7, LoTm, Salt, Sw, Gly, Lbstat
S1192	new_3213597_3213853_c	3213597	3213853	intra	*NA*	SigB	Etha, Sw, Salt, Gly, BC, M9-stat, Lbstat, Heat, LPhT, S0
S1193	new_3217011_3217663	3217011	3217663	Indep-NT	*yugS*	SigB	Etha, Sw, Gly, Salt, BC, B60, Heat, S8, B36, HiTm
S1199	new_3224752_3224810	3224752	3224810	5'	*NA*	SigB	Etha, Salt, Heat, M9-stat, LPhT, Sw, Gly, BC, HiTm, B36
S1202	new_3225687_3226671	3225687	3226671	Indep-NT	*yugH*	SigB	LoTm, C30, Salt, S7, S5, S8, Sw, S3, C90, S6

Feature name[a]	Locus tag[a]	Start position[a]	End position[a]	Classification[a,b]	Antisense name[a]	TU short sig[a]	Highly expressed conditions[a,c]
S1224	new_3266637_3269004	3266637	3269004	Indep-NT	yueB	SigB	Etha, Gly, Sw, Salt, LoTm, C30, BT, HiTm, Mal, G135
S1225	new_3273703_3274665	3273703	3274665	Indep-NT	yukB	SigB	Etha, Sw, Gly, Salt, B60, M9-stat, LoTm, T0.30H, BT, LPhT
S1243	new_3307598_3308367	3307598	3308367	indep	yutK	SigB	Etha, Salt, Gly, M9-stat, Sw, LoTm, B60, BC, Heat, S0
S1244	new_3308369_3308785	3308369	3308785	inter	yuzB	SigB	Salt, LoTm, Gly, Etha, Sw, Heat, S0, C30, M9-stat, Mal
S1276	new_3404438_3404707_c	3404438	3404707	5'	NA	SigA, SigB	LBGtran, LBGexp, Etha, LPh, S3, Lbexp, G180, dia0, HPh, Mt0
S1277	new_3404709_3405636_c	3404709	3405636	inter	yvrD	SigB	Etha, B60, Sw, M9-stat, B36, C30, C90, Salt, G135, Gly
S1283	new_3443530_3443612_c	3443530	3443612	intra	NA	SigA, SigB	Diami, dia5, dia15, Gly, T0.30H, Heat, T1.0H, T1.30H, T0.0H, LPhT
S1287	new_3451771_3451862_c	3451771	3451862	inter	NA	SigB, SigA	Etha, S8, Cold, S7, G135, G180, G150, H2O2, M9-stat, S6
S1290	new_3460206_3462957	3460206	3462957	Indep-NT	opuBC	SigB	Sw, M9-stat, B60, Salt, LoTm, T0.30H, BC, Gly, B36, T1.0H
S1318	new_3546111_3546186	3546111	3546186	5'	NA	SigB	Salt, S3, S4, Etha, S5, LoTm, Gly, M9-stat, BT, S6
S1319	new_3546861_3547259	3546861	3547259	3'PT	pgcM	SigB	Etha, dia15, Heat, LoTm, M9-stat, Gly, dia5, Diami, S8, H2O2
S1326	new_3574321_3577707	3574321	3577707	indep	yvcD	SigB	Sw, M9-stat, B60, T0.30H, Gly, BC, B36, T1.0H, Salt, C30
S1341	new_3609903_3610036_c	3609903	3610036	inter	NA	SigA, SigB	H2O2, Sw, M40t45, M40t90, Diami, Salt, dia15, BC, G135, C90
S1342	new_3614978_3615063_c	3614978	3615063	inter	NA	SigA, SigB	Gly, Sw, B60, BT, LoTm, Salt, Etha, M9-stat, B36, S0
S1343	new_3615385_3615449_c	3615385	3615449	5'	NA	SigB	Gly, M9-stat, B60, Sw, Etha, LoTm, BT, B36, T1.30H, T2.0H
S1358	new_3647109_3647701_c	3647109	3647701	Indep-NT	yvyE	SigB	Etha, M9-stat, Gly, C30, B60, Salt, T0.30H, Sw, LoTm, S8
S1367	new_3669959_3672791	3669959	3672791	5'	ggaA	SigB	Sw, LoTm, M9-stat, Salt, B60, T0.30H, Gly, T1.0H, B36, BI
S1370	new_3672793_3672923	3672793	3672923	5'	NA	SigA, SigB	Sw, T0.30H, LoTm, B60, T1.0H, Lbtran, C30, LBGtran, dia5, dia15

Feature name[a]	Locus tag[a]	Start position[a]	End position[a]	Classification[a,b]	Antisense name[a]	TU short sig[a]	Highly expressed conditions[a,c]
S123	new_368964_369186	368964	369186	5'	NA	SigB	Etha, Salt, B60, Sw, B36, Gly, Heat, S6, Cold, S7
S1378	new_3691599_3692493_c	3691599	3692493	3'PT	gerBC	SigB	Etha, Sw, Salt, Heat, S8, B36, Gly, S7, LPhT, S6
S1379	new_3693907_3693970_c	3693907	3693970	5'	NA	SigB	Etha, Salt, Gly, Heat, Sw, LPhT, HiTm, Diami, B60, UNK1
S1382	new_3699152_3701361	3699152	3701361	5'	pgsB	SigB	Sw, Salt, dia5, BC, B60, Diami, Lbstat, H2O2, dia15, Gly
S1383	new_3707708_3707818	3707708	3707818	inter	NA	SigB	Etha, M9-stat, Sw, M40t90, G150, B60, T0.30H, M0t90, T1.30H, Lbexp
S1384	new_3708133_3708785	3708133	3708785	3'UTR	ywrO	SigB	Etha, M9-stat, S5, S7, S6, Sw, S8, B60, G150, S4
S1409	new_3762479_3762776_c	3762479	3762776	5'	clsA	SigB, SigEF	Etha, S6, S3, S4, S8, S5, S7, M9-stat, Sw, Salt
S1413	new_3769109_3769193_c	3769109	3769193	5'	NA	Sig-, SigH, SigB	BT, B60, B36, S1, Etha, LoTm, M9-stat, S2, T0.30H, BC
S1414	new_3769195_3769288_c	3769195	3769288	inter	NA	SigH, SigB	BT, Etha, B60, S1, B36, BC, S2, LoTm, T3.0H, T3.30H
S1415	new_3769454_3769625_c	3769454	3769625	intra	NA	SigA, SigB	Etha, BT, S1, GM-0.2, GM-0.1, Salt, GM+5, T-5.40H, MG+60, MG+120
S1416	new_3769809_3769955_c	3769809	3769955	5'	NA	SigA, SigB	Etha, BT, S1, Salt, GM-0.2, GM-0.1, M9-stat, GM+5, GM+45, UNK1
S1417	new_3769957_3770070_c	3769957	3770070	inter	NA	SigB	Etha, Salt, M9-stat, S8, B60, B36, Sw, Gly, LoTm, C30
S1418	new_3774188_3774399_c	3774188	3774399	inter	NA	SigB	Etha, Sw, Salt, Heat, S8, Gly, S6, S7, BC, S5
S1446	new_3819213_3819619	3819213	3819619	3'PT	ywjB	SigB	Etha, C30, Sw, Gly, C90, M/G, BC, Salt, Fru, G135
S1459	new_3873771_3874382	3873771	3874382	Indep-NT	bacA	SigB	Etha, Salt, Gly, Heat, C30, LPhT, HiTm, C90, S8, M/G
S1477	new_3943322_3944525_c	3943322	3944525	3'PT	sacY	SigB	Etha, B36, B60, M9-stat, Salt, Sw, LoTm, C30, Gly, Heat
S1485	new_3971340_3971882_c	3971340	3971882	Indep-NT	yxlA	SigB	Etha, LBGstat, T1.0H, T0.30H, Diami, T1.30H, dia15, G150, G135, G180
S1496	new_3996629_3996729	3996629	3996729	5'	NA	SigB, SigWXY	Etha, Gly, Sw, Salt, B60, T0.30H, LoTm, M9-stat, T1.0H, B36

Feature name[a]	Locus tag[a]	Start position[a]	End position[a]	Classification[a,b]	Antisense name[a]	TU short sig[a]	Highly expressed conditions[a,c]
S1497	new_3996731_3996828	3996731	3996828	5'	*NA*	SigD, SigB, SigWXY	Etha, T0.30H, Gly, T1.0H, Sw, Salt, T1.30H, GM+15, B60, GM-0.1
S1519	new_4053373_4053453_c	4053373	4053453	inter	*NA*	SigB	Etha, Diami, Salt, MG+10, Gly, MG+60, G180, GM+60, G135, dia5
S1539	new_4109864_4110916_c	4109864	4110916	3'PT	*yxaC*	SigB	Etha, C30, H2O2, Salt, Paraq, GM+120, Diami, Oxctl, G180, Sw
S1556	new_4159220_4159541	4159220	4159541	inter	*yyzB*	SigB	Etha, C30, Salt, S6, S8, M9-stat, S7, Oxctl, Paraq, Heat
S1560	new_4166589_4166725_c	4166589	4166725	5'	*NA*	SigA, SigB	H2O2, G150, G180, Cold, Paraq, G135, Oxctl, Salt, Etha, BMM
S139	new_469286_469367	469286	469367	5'	*NA*	SigB	Etha, S5, S6, ferm, S8, nit, S7, BT, Salt, Sw
S140	new_474268_474634	474268	474634	indep	*NA*	SigA, SigB, SigA	Etha, Salt, Cold, Sw, B60, Gly, M9-stat, LPhT, B36, G135
S141	new_478812_478943	478812	478943	5'	*NA*	SigB, SigA	Etha, Salt, Heat, Cold, H2O2, G180, LPhT, Oxctl, Paraq, G135
S145	new_490639_490854	490639	490854	Indep-NT	*ydzK*	SigEF, SigB	Etha, S4, S8, S5, C30, S7, B60, S6, Salt, B36
S147	new_492500_492577_c	492500	492577	5'	*NA*	SigK, SigB	S6, S8, S7, Etha, S5, Salt, BT, Heat, B36, B60
S148	new_495406_495699_c	495406	495699	3'PT	*ydbC*	SigB	Etha, Salt, M9-stat, C30, Heat, Sw, B60, S8, C90, UNK1
S163	new_519083_519559_c	519083	519559	indep	*ndoA*	SigD, SigB, SigK	Etha, Gly, Sw, S8, C30, S6, LBGstat, T0.30H, S7, Salt
S164	new_519561_520160_c	519561	520160	5'	*rsbRA*	SigB, SigK	Etha, Sw, Gly, S6, S8, S7, B60, Salt, BT, M9-stat
S190	new_598065_598153	598065	598153	5'	*NA*	SigK, SigB, SigA	Etha, BT, S6, S7, S8, G/S, GM+10, Gly, GM+15, B36
S34	new_60361_60429	60361	60429	intra	*NA*	Sig-, SigB, SigA	Etha, S5, S4, S7, S8, S6, S3, H2O2, Diami, G180
S35	new_64001_64098	64001	64098	5'	*NA*	SigGF, Sig-, SigB, SigA	S5, S7, S8, S6, S4, Etha, S3, H2O2, G135, G180
S230	new_697399_697485	697399	697485	5'	*NA*	SigB	Etha, Gly, M9-stat, Sw, B60, S4, S6, LPhT, S3, S8

Feature name[a]	Locus tag[a]	Start position[a]	End position[a]	Classification[a,b]	Antisense name[a]	TU short sig[a]	Highly expressed conditions[a,c]
S236	new_728319_728449_c	728319	728449	5'	NA	SigB	Salt, HiOs, Etha, Gly, MG+25, MG+45, M9-exp, MG+t5, MG+15, GM+5
S261	new_827843_827923	827843	827923	inter	NA	SigB	Etha, Salt, B36, LoTm, Heat, B60, M9-stat, S8, C30, LPhT
S269	new_843132_844096	843132	844096	5'	yflE	SigB	Etha, Salt, Sw, Gly, C30, M9-stat, S8, B60, BC, S7
S272	new_846186_849552	846186	849552	3'NT	yfkR	SigB	Salt, Sw, Etha, LoTm, B36, Heat, LPhT, BC, T0.30H, H2O2
S273	new_860294_861944	860294	861944	inter	yfkL	SigB	Etha, S6, S8, S5, S7, Gly, Salt, LoTm, S4, Diami
S275	new_863681_863837	863681	863837	indep	NA	SigGF, SigA, SigB	S8, Etha, S6, Salt, S7, S5, ferm, nit, BT, Sw

[a]The feature name, locus tag name, start and end positions, their classification were extracted from Nicolas *et al.,* 2012, features are sorted based on the locus tag

[b] The descriptions of classification were mentioned in the legend of Fig. 13

[c] Condition full descriptions were mentioned in Table 1 (for details, see Materials and Methods)

Appendix 3.2.2: Motifs and their scores for the newly identified genes (90 genes) along with classical SigB genes (40 genes)

Gene[a]	BSU number	Comment	-35 box[b]	-10 box[b]	-35 box score[c]	-10 box score[c]	Motif score
aldY	BSU38830	Newly identified SigB	GTTTAC	GGAAAT	4.84	4.55	9.39
atpC	BSU36800	Newly identified SigB	GTTTAA	GGGTAA	5.83	5.67	11.50
cdd	BSU25300	Newly identified SigB	GTTCAA	GGGTAA	0.89	5.67	6.56
cpgA	BSU15780	Newly identified SigB	GATTGA	GGGTAC	1.88	3.61	5.49
cypC	BSU02100	Newly identified SigB	GTTTAT	GTGAAA	5.64	1.51	7.15
dgkA	BSU25310	Newly identified SigB	GTTCAA	GGGTAA	0.89	5.67	6.56
drm	BSU23500	Newly identified SigB	GTTTAT	GGGAAA	5.64	5.84	11.48
eag	BSU13650	Newly identified SigB	GTTTAA	GGGTAT	5.83	5.81	11.64
era	BSU25290	Newly identified SigB	GTTCAA	GGGTAA	0.89	5.67	6.56
folD	BSU24310	Newly identified SigB	GTTTCG	GGGTAT	2.77	5.81	8.58
galE	BSU38860	Newly identified SigB	GTTTAG	CGAGAA	4.50	-0.92	3.57
galK	BSU38200	Newly identified SigB	GTTTTC	GGGAAT	4.58	5.98	10.56
galT	BSU38190	Newly identified SigB	GTTTTC	GGGAAT	4.58	5.98	10.56
glpD	BSU09300	Newly identified SigB	GTTTCA	GGGAAG	4.11	4.29	8.40
hemA	BSU28170	Newly identified SigB	GTTTGA	GGGAAA	4.04	5.84	9.88
ipi	BSU11130	Newly identified SigB	GTTCAA	GGGAAA	0.89	5.84	6.74
ispA	BSU24280	Newly identified SigB	GTTTCG	GGGTAT	2.77	5.81	8.58
lytD	BSU35780	Newly identified SigB	GTTTGT	GGGAAA	3.85	5.84	9.69
malS	BSU29880	Newly identified SigB	GTTTAC	GGCTAA	4.84	1.63	6.48
menC	BSU30780	Newly identified SigB	GTGTCG	GGGTAT	0.76	5.81	6.57
nagBA	BSU35020	Newly identified SigB	GTTTAC	GGATAT	4.84	4.37	9.22
nsrR	BSU09380	Newly identified SigB	GTTTGC	GGGTAA	3.05	5.67	8.72
nusB	BSU24320	Newly identified SigB	GTTTCG	GGGTAT	2.77	5.81	8.58
opuD	BSU30070	Newly identified SigB	GTTTTG	GGGTAA	4.23	5.67	9.90
pgcA	BSU09310	Newly identified SigB	GTTTAT	GGGTAA	5.64	5.67	11.30
phoH	BSU25340	Newly identified SigB	GTTCAA	GGGTAA	0.89	5.67	6.56
pupG	BSU23490	Newly identified SigB	GTTTAT	GGGAAA	5.64	5.84	11.48
recO	BSU25280	Newly identified SigB	GTTCAA	GGGTAA	0.89	5.67	6.56
ripX	BSU23510	Newly identified SigB	GTTTAT	GGGAAA	5.64	5.84	11.48
rpe	BSU15790	Newly identified SigB	GATTGA	GGGTAC	1.88	3.61	5.49
rsoA	BSU33222	Newly identified SigB	GTTAGC	GGGAAA	-0.93	5.84	4.91
spo0E	BSU13640	Newly identified SigB	GTTTAA	GGGTAT	5.83	5.81	11.64
spoVM	BSU15810	Newly identified SigB	GATTGA	GGGTAC	1.88	3.61	5.49
tcyA	BSU03610	Newly identified SigB	GTTTGA	AGGTAC	4.04	2.40	6.44
tcyB	BSU03600	Newly identified SigB	GTTTGA	AGGTAC	4.04	2.40	6.44
tcyC	BSU03590	Newly identified SigB	GTTTGA	AGGTAC	4.04	2.40	6.44
thiN	BSU15800	Newly identified SigB	GATTGA	GGGTAC	1.88	3.61	5.49
tmk	BSU00280	Newly identified SigB	GCATGA	GGGAAT	0.21	5.98	6.19

Gene[a]	BSU number	Comment	-35 box[b]	-10 box[b]	-35 box score[c]	-10 box score[c]	Motif score
ung	BSU37970	Newly identified SigB	GTTTTC	GGGTAT	4.58	5.81	10.39
xseA	BSU24300	Newly identified SigB	GTTTCG	GGGTAT	2.77	5.81	8.58
xseB	BSU24290	Newly identified SigB	GTTTCG	GGGTAT	2.77	5.81	8.58
yabS	BSU00650	Newly identified SigB	GGTTAT	GGGAAA	3.47	5.84	9.31
yckC	BSU03390	Newly identified SigB	GTTTTC	GTGTAT	4.58	1.48	6.05
yczO	BSU04039	Newly identified SigB	GTTTGA	AGGAAA	4.04	4.63	8.67
ydfO	BSU05490	Newly identified SigB	GTTTCG	GGGGAG	2.77	2.13	4.91
ydfP	BSU05500	Newly identified SigB	GTTTCG	GGGGAG	2.77	2.13	4.91
yebE	BSU06400	Newly identified SigB	TTTTGA	GGGTAA	2.31	5.67	7.98
yebG	BSU06410	Newly identified SigB	TTTTGA	GGGTAA	2.31	5.67	7.98
yetO	BSU07250	Newly identified SigB	GAATAA	GGGAAA	1.93	5.84	7.77
yflB	BSU07735	Newly identified SigB	GTTTTA	TGGAAA	5.56	3.52	9.08
yflD	BSU07720	Newly identified SigB	GTTTTA	TGGAAA	5.56	3.52	9.08
yhbJ	BSU09000	Newly identified SigB	GATTCA	GGGAAG	1.94	4.29	6.23
yhcA	BSU09010	Newly identified SigB	GATTCA	GGGAAG	1.94	4.29	6.23
yhcB	BSU09020	Newly identified SigB	GATTCA	GGGAAG	1.94	4.29	6.23
yhcC	BSU09030	Newly identified SigB	GATTCA	GGGAAG	1.94	4.29	6.23
yhcY	BSU09320	Newly identified SigB	GTTTAT	GGGTAA	5.64	5.67	11.30
yhcZ	BSU09330	Newly identified SigB	GTTTAT	GGGTAA	5.64	5.67	11.30
yhgE	BSU10160	Newly identified SigB	GATTGG	CAGGAA	0.54	-2.79	-2.26
yitT	BSU11120	Newly identified SigB	GTTCA	GGGAAA	0.70	5.84	6.54
yjzE	BSU11839	Newly identified SigB	GTTTCA	GGGAAA	4.11	5.84	9.95
ykzQ	BSU13789	Newly identified SigB	GATTTC	GGGAAA	2.41	5.84	8.25
yojJ	BSU19430	Newly identified SigB	GTTTCG	GGGCAT	2.77	3.48	6.26
yqbM	BSU26060	Newly identified SigB	AATTAC	GGGTAT	0.22	5.81	6.03
yqeF	BSU25700	Newly identified SigB	GTTTTT	AGGGAA	5.37	2.48	7.84
yqfF	BSU25330	Newly identified SigB	GTTCAA	GGGTAA	5.37	5.67	11.04
yqfG	BSU25320	Newly identified SigB	GTTCAA	GGGTAA	0.89	5.67	6.56
yqgC	BSU25030	Newly identified SigB	GTTTGA	GGGAAA	4.04	5.84	9.88
yqhY	BSU24330	Newly identified SigB	GTTTCG	TGGGTA	2.77	-1.23	1.54
yqjF	BSU23900	Newly identified SigB	GTTTAA	GGGAAT	5.83	5.98	11.81
yqzK	BSU23519	Newly identified SigB	GTTTAT	GGGAAA	5.64	5.84	11.48
yqzL	BSU25289	Newly identified SigB	GTTCAA	GGGTAA	0.89	5.67	6.56
yrhK	BSU27150	Newly identified SigB	GTTTAC	GGGTAA	4.84	5.67	10.51
ytpI	BSU29260	Newly identified SigB	GTTTCA	AGGGAA	4.11	2.48	6.59
ytrP	BSU29650	Newly identified SigB	GTTTAC	AGGAAA	4.84	4.63	9.47
yugU	BSU31280	Newly identified SigB	GTTTAT	AGGAAA	5.64	4.63	10.26
yutJ	BSU32200	Newly identified SigB	GTTTGC	GGGTAT	3.05	5.81	8.86
yuzH	BSU31279	Newly identified SigB	GTTTAT	AGGAAA	5.64	4.63	10.26
yvoA	BSU35030	Newly identified SigB	GTTTAC	GGATAT	4.84	4.37	9.22
yvyI	BSU35790	Newly identified SigB	GTTTGT	GGGAAA	3.85	5.84	9.69

Gene[a]	BSU number	Comment	-35 box[b]	-10 box[b]	-35 box score[c]	-10 box score[c]	Motif score
yvzE	BSU35699	Newly identified SigB	GTTTAT	GGGTAA	5.64	5.67	11.30
ywkB	BSU37040	Newly identified SigB	GAGTAA	GGGAAA	1.65	5.84	7.49
ywlB	BSU36960	Newly identified SigB	GAATAA	GGGTAA	1.93	5.67	7.60
ywmF	BSU36680	Newly identified SigB	GTTTAT	AGGAAA	5.64	4.63	10.26
ywnF	BSU36580	Newly identified SigB	GTATAA	GGGAAT	4.10	5.98	10.08
yxkA	BSU38870	Newly identified SigB	GTTTAG	CGAGAA	4.50	-0.92	3.57
yybS	BSU40520	Newly identified SigB	GTTTTG	GGGAAA	4.23	5.84	10.07
yybT	BSU40510	Newly identified SigB	GTTTTG	GGGAAA	4.23	5.84	10.07
yycR	BSU40250	Newly identified SigB	GAAGTT	GGGTAT	-3.47	5.81	2.34
yyzG	BSU40259	Newly identified SigB	GAAGTT	GGGTAT	-3.47	5.81	2.34
yyzH	BSU40529	Newly identified SigB	GTTTTG	GGGAAA	4.23	5.84	10.07
bmr	BSU24010	Classical SigB	GTTTAC	GGGGAT	4.84	3.83	8.67
bmrR	BSU24020	Classical SigB	GTTTAC	GGGGAT	4.84	3.83	8.67
bmrU	BSU24000	Classical SigB	GTTTAC	GGGGAT	4.84	3.83	8.67
bofC	BSU27750	Classical SigB	GATTAC	GGGTAT	2.68	5.81	8.49
clpC	BSU00860	Classical SigB	GTTTTG	TGGAAA	4.23	3.52	7.74
clpP	BSU34540	Classical SigB	GTTTGA	GGGGAA	4.04	3.69	7.73
csbA	BSU35180	Classical SigB	GATTGA	GGGTAT	1.88	5.81	7.69
csbB	BSU08600	Classical SigB	GTTTAA	AGGAAT	5.83	4.77	10.60
csbC	BSU39810	Classical SigB	GTTTCA	GGGTAC	4.11	3.61	7.72
csbD	BSU36670	Classical SigB	GTTTAT	GGGAAG	5.64	4.29	9.92
csbX	BSU27760	Classical SigB	GATTAC	GGGTAT	2.68	5.81	8.49
ctc	BSU00520	Classical SigB	GTTTAA	GGGTAT	5.83	5.81	11.64
ctsR	BSU00830	Classical SigB	GTTTTG	TGGAAA	4.23	3.52	7.74
dps	BSU30650	Classical SigB	GTTTTA	GGGTAT	5.56	5.81	11.37
gabD	BSU03910	Classical SigB	GATTAC	GGGAAT	2.68	5.98	8.66
gsiB	BSU04400	Classical SigB	GTTTAA	GGGAAT	5.83	5.98	11.81
gspA	BSU38430	Classical SigB	GTGTTT	GGGTAT	3.35	5.81	9.16
katE	BSU39050	Classical SigB	GTTTAT	GGGTAA	5.64	5.67	11.30
katX	BSU38630	Classical SigB	GTTTTA	GGGAAT	5.56	5.98	11.55
mcsA	BSU00840	Classical SigB	GTTTTG	TGGAAA	4.23	3.52	7.74
mcsB	BSU00850	Classical SigB	GTTTTG	TGGAAA	4.23	3.52	7.74
rsbV	BSU04710	Classical SigB	GTTTAA	GGGTAT	5.83	5.81	11.64
rsbW	BSU04720	Classical SigB	GTTTAA	GGGTAT	5.83	5.81	11.64
rsbX	BSU04740	Classical SigB	GTTTAA	GGGTAT	5.83	5.81	11.64
sigB	BSU04730	Classical SigB	GTTTAA	GGGTAT	5.83	5.81	11.64
trxA	BSU28500	Classical SigB	GTTTTA	GGGCAT	5.56	3.48	9.05
ydaD	BSU04190	Classical SigB	GTTTAT	AGGTAC	5.64	2.40	8.04
ydaE	BSU04200	Classical SigB	GTTTAT	AGGTAC	5.64	2.40	8.04
ydaG	BSU04220	Classical SigB	GTTTAA	TGGAAA	5.83	3.52	9.35
ydaP	BSU04340	Classical SigB	GTTTTA	TGGTAT	5.56	3.49	9.05
yflA	BSU07750	Classical SigB	GTTTAT	GGGAAA	5.64	5.84	11.48
ytxG	BSU29780	Classical SigB	GTTTAT	GGGTAA	5.64	5.67	11.30

Gene[a]	BSU number	Comment	-35 box[b]	-10 box[b]	-35 box score[c]	-10 box score[c]	Motif score
ytxH	BSU29770	Classical SigB	GTTTAT	GGGTAA	5.64	5.67	11.30
ytxJ	BSU29760	Classical SigB	GTTTAT	GGGTAA	5.64	5.67	11.30

[a] Genes are sorted based on gene name
[b] Motifs are extracted from the supplementary table S4 of Nicolas *et al.*, 2012
[c] Motif scores are calculated according to the method of Staden *et al.* (for details, see Materials and Methods)