Chapter 8

General Summarizing Discussion

The main common theme of the research described in this thesis is the regulation of cellular processes in the bacterium *Bacillus subtilis* at the level of RNA. In the founding years of molecular biology, RNA was viewed solely as the messenger between the genetic information in the DNA and the proteins. In the last two decades, however, it has become increasingly clear that RNA is very important at the level of post-transcriptional regulation as well (1). Studies on RNA have been somewhat complicated by its instability and its relatively low abundance in the cell. However, since the early days of molecular biology, enormous technological advances have been achieved, which made the research described in this thesis possible. This progress is best illustrated by the sequencing of the first RNA molecule, alanine transfer RNA (tRNA) from baker's yeast in the 1960s, which took several years and involved a staggering 140 kg of yeast cells (2, 3). Nevertheless, this enormous effort led to the first sequence of an RNA molecule and the identification of its modified bases. Based on this sequence, three possible secondary structures were proposed for this RNA, one of which is the now iconic tRNA cloverleaf structure. It was this structure that inspired the artwork on the cover of this thesis. The tRNA cloverleaf structure is also the first example of a sequence-structure relationship and it led Crick to propose the Wobble base theory of codon recognition (4). Earlier, in the late 1950s, Holley and others had already implicated RNA in the process of amino acid charging based on their observation that this process was dependent on ribonuclease (2). Subsequently, the respective molecule was identified as tRNA.

Evaluation of experimental approaches to probe the B. subtilis transcriptome

Today, many advanced technologies are available for the detailed analysis of complete transcriptomes. Notably, such transcriptome analyses are generally performed with reverse-transcribed RNA. While this has the advantage that the challenges of directly sequencing RNA are circumvented, the downside of this approach is that potentially modified nucleotides will be missed (5). In the most advanced scenario, the resulting copy DNA (cDNA) is directly sequenced (RNA-seq) but, until very recently, the cDNA was mostly annealed to micro- or tiling arrays. The latter tiling array-based approach was applied for a transcript profiling analysis of *B. subtilis* across 104 different conditions as described in **Chapter 2** of this thesis, and for studies on a small regulatory RNA (srRNA) mutant described in **Chapter 4**. The transcriptome data thus obtained were first used to characterize the overall regulatory architecture of *B. subtilis* (**Chapter 2**), and subsequently to identify the most likely interactions of a small regulatory RNA molecule (srRNA) with messenger RNA (mRNA) molecules out of a large set of predicted interactions (**Chapter 4**). This focus on transcriptomics as an experimental approach to investigate global gene regulation in a bacterium raises the question how relevant the determined RNA expression levels actually are for the study of RNA functions. The answer to the first question seems to be that it can be sufficient to focus on expression levels alone in studies on RNA function, but that it will depend on the regulatory mechanism in which the respective RNA molecule is involved whether this approach gives meaningful results. For instance, as described in **Chapter 6,** a negative expression correlation and the genomic organization were the only indications for the S1136-S1134 antisense RNA (asRNA)-mediated downregulation of the *rpsD* gene expression, and this was subsequently experimentally confirmed. As was also noted in **Chapter 3**, such negative expression correlations between sense and antisense transcripts seem to be an important indicator of asRNA regulation. Conversely, in **Chapter 4** it was reported that expression correlations between predicted srRNA-mRNA pairs are not stronger than those of essentially random pairs. This shows that, for srRNA, expression correlations are by themselves no indication of target regulation. For this reason, more focused expression correlation analyses, namely Peaks correlations and B-cluster enrichment, were performed. Since these analyses are more focused, the srRNA-mRNA pairs derived from these analyses may be more relevant than

random pairs. The extent of expression correlation between an srRNA and its mRNA target most likely depends on the involved regulatory mechanism. Regulatory mechanisms of srRNAs can be divided into mechanisms that trigger mRNA degradation and mechanisms that inhibit translation. Both of these will most likely result in some change in the target mRNA level, but this would be more pronounced in the first than the latter. It would be interesting to see if, indeed, those mRNA degradation-triggering srRNA-target interactions are the ones preferentially surfacing in the employed correlation analyses. However, it is impossible to say this now since only very few srRNA-mRNA interactions have been confirmed for *B. subtilis*. A general downside of using transcriptomic analyses for identifying regulatory RNA interactions is that, most likely, not all regulation will be detectable at the RNA level. An alternative strategy would therefore be to measure the abundances of proteins through proteomics approaches. Two-dimensional (2D) PAGE-based analyses with subsequent identification of protein spots with different intensities has been employed for the studies described in **Chapter 4** on a FsrA/S512 srRNA mutant, and this identified multiple differentially expressed proteins in the stationary growth phase on Lysogeny Broth (LB) medium. Since proteins are the ultimate physiological effectors of cellular traits, proteomics seems the most inclusive method for identifying post-transcriptional RNA regulation. This is especially so when quantitative proteomics can be employed. However, proteomic analyses are technically challenging and lack genome-wide coverage. These experimental limitations of proteomics only allow a subset of the proteins of an organism to be analyzed, depending on the particular analysis that is being performed. In the experiments described in **Chapter 4** for instance only those cytosolic proteins with a PI between 4 and 7 could be identified. In contrast, transcriptome analyses do provide genome-wide coverage, although they can be technically challenging as well. Valid for both transcriptomics and proteomics is that only genes and proteins that are expressed in the experimental condition can be identified. Because of this, it is important to study regulatory RNAs in their natural context, for instance in the condition where a phenotype has been identified for a regulatory RNA mutant. Because it is not realistic to obtain experimental evidence in every possible environmental condition, this implies that, to obtain a comprehensive view of srRNA regulation, one has to rely to some extent on target predictions (6).

Once a putative srRNA-mRNA interaction has been identified either by transcriptomics, proteomics or target predictions, the direct interaction between the two RNA molecules has to be confirmed. The state of the art for this is the construction of one or multiple point mutations in the one RNA molecule (e.g. the srRNA) to abolish regulation and compensatory mutations in the other RNA molecule (e.g. the mRNA) to restore this regulation. An example of this is presented in **Chapter 5** for the RnaC/S1022-*abrB* interaction. This represents actually the first report for *B. subtilis* that a single nucleotide substitution can be sufficient to abolish, and subsequently restore, an srRNA-mRNA interaction. One could argue that compensatory mutations are still no absolute proof of a direct srRNA-mRNA interaction, but such genetic evidence is currently the only possible way for obtaining *in vivo* experimental evidence for direct srRNA-mRNA interactions. The observation that a single nucleotide substitution can suffice to abolish an srRNA-mRNA interaction also has implications for both the evolution and prediction of srRNA-mRNA interactions. Firstly, **Chapter 3** presents some comments on the evolution of regulatory RNA-mediated regulation of gene expression. It is noted that very slight sequence alterations can be sufficient to induce srRNA expression or confer srRNA regulation. This view is directly supported by the observed importance of a single nucleotide in RnaC/S1022 for *abrB* regulation. The respective nucleotide is the middle nucleotide of the fourth codon of the *abrB* gene. Since mutation of this nucleotide will also change the *abrB* coding sequence, the constraint on mutations of this nucleotide might "lock" the RnaC/S1022 interaction in evolution. Regulation via the middle nucleotide of a codon is distinct from an interesting observation in *Salmonella*

where a wobble base pair was essential for distinction of two paralogous gene targets by the SgrS srRNA (7).

The green fluorescent protein (GFP) has been repeatedly used as a reporter for promoter activity or protein expression in the research described throughout this thesis. In the studies for **Chapter 4**, GFP was translationally coupled to a possible srRNA target of the RsaE/S415 srRNA, and the respective hybrid gene was placed on a plasmid to verify this target's deregulation in an RsaE/S415 deletion mutant. In the same chapter the *gfp* gene was integrated in the genome downstream of the promoter of RsaE/S415, S462 or FsrA/S512 to obtain information on the expression of these (putative) srRNAs. Similar promoter fusions were constructed in the studies documented in **Chapter 5** for *hag*, RnaC/S1022 and *abrB*. Furthermore, a chromosomally-located translational fusion of AbrB and GFP was employed to get insights into the cell-to-cell variation of the AbrB protein. Finally, in **Chapter 7** a plasmid-located *gfp* promoter fusion is described that was employed to facilitate the construction of a series of point mutations in the *bmrB* regulatory element to unravel its regulatory mechanism. Together the respective experiments illustrate the possibilities and ease of use of GFP reporter constructs. However, as with every experimental technique the possible influence of the reporter construct on the experimental outcome has to be taken into account. Attaching a GFP molecule to a protein can for instance have an influence of the stability of the respective protein. This might in turn influence the experimental outcome, for instance in the studies described in **Chapter 5**. Another consequence of the use of translationally-coupled reporter constructs is that they are often truncated to only include the predicted srRNA interaction region. However, such truncated constructs may exclude sequence regions that are crucial for regulation. These could for instance involve downstream RNase processing sites, as was shown for the RyhB-*sodB* interaction in *E. coli* (8). In addition, and as was also discussed in **Chapter 5**, it is important to place reporter constructs as much as possible in their native chromosomal context, because the chromosomal location can influence the experimental outcome. Such considerations seem especially important for the analysis and interpretation of subtle phenotypes, such as the noise generation presented in **Chapter 5**.

The search for srRNA targets can be approached from two perspectives, namely with a focus on an srRNA mutant phenotype, or with a focus on deregulated targets. With the first focus, phenotypical tests can be performed on an srRNA mutant, for instance in a condition where the srRNA is expressed to a high level. Alternatively, the srRNA can be overexpressed to inspect growth effects. Subsequently, when a phenotype is observed one can try to explain this phenotype via target predictions and targeted experiments. This is the approach used for the present PhD thesis research, as illustrated in **Chapters 4** and **5**. An alternative to this is the direct experimental identification of an srRNA-target interaction. Its biological function and physiological relevance can then be explained later. This is best approached via so-called pulsed overexpression of the srRNA from a plasmid in the relevant growth condition rather than by relying on an srRNA mutant strain. One can subsequently compare the RNA levels (e.g. via RNA-seq) in the strain in which the srRNA was induced for a short period of time to the RNA levels in a strain where a mock RNA was induced or which carries the empty plasmid vector. The main advantage of this pulsed overexpression approach is that indirect effects resulting from the absence of the srRNA are expected to be largely absent. If for instance an important transcription factor is an srRNA target, its complete regulon will be deregulated in the deletion mutant approach, but not in the pulsed overexpression approach. The latter approach was pioneered by the group of Jörg Vogel for Hfq-binding srRNAs in *E. coli* and *Salmonella*, and it is currently being established for *B. subtilis* as well (E. L. Denham, personal communication). Notably, RNA-seq has also been adapted to identify processed RNAs and this is expected to reveal much about the role of RNA processing in RNA regulation (9). An FsrA/S512 mutant

was used for proteomic and transcriptomic analyses as described in **Chapter 4** and it is clear from the subsequent analyses that many of the identified deregulated proteins or genes are the result of indirect effects. However, the approach of studying srRNA deletion mutants to identify srRNA targets has also proven successful as illustrated by the studies described in **Chapter 4** and especially in **Chapter 5**. It will thus depend on the ease of finding a phenotype for an srRNA mutant and the further goals of a particular project to decide which approach can best be taken for a detailed functional srRNA characterization.

Other extensively used approaches for the research described in this thesis are computational of nature. Computational and biostatistical analyses were used to completely annotate the regulatory network of *B. subtilis* based on extensive tiling array data across a large number of conditions (**Chapter 2**), for analysis of the evolutionary conservation and secondary structure of almost 1600 RNA segments (**Chapter 3**), for prediction of srRNA functions and the most likely srRNA targets (**Chapter 4**), and to stochastically model srRNA regulation dynamics and its accompanying noise-inducing aspects (**Chapter 5**). It is clear that the research described in this thesis would not have been possible without the enormous advantages provided by computational methods for the analysis of biological data. It also shows that, whenever bioinformaticians manage to effectively communicate with biologist and *vice-versa*, many new principles can be discovered at the interface of (systems) biology and computational biology. The only downside of computational predictions is that predictions are "just predictions". They are often required, but their value is questionable until suitable biological evidence has been obtained to confirm these predictions. Rigorous biological verification of bioinformatics predictions was the initial goal of the studies reported in **Chapter 4**, but the amount of work this would entail was not feasible within the time frame of the present PhD thesis research project. For this reason, the value of target predictions was instead illustrated with the already published data on the FsrA/S512, SR1, S462, and RsaE/S415 srRNAs described in **Chapter 4** and on RnaC/S1022 described in **Chapter 5**.

In the next paragraphs of the present chapter, the main discussion points that have arisen from the different studies presented in this thesis will be discussed.

New putative regulatory RNAs in B. subtilis

Chapter 2 reports that the expression space of *B. subtilis* is highly variable and condition-dependent. This is illustrated by the observations that a very small percentage of genes (4.4%) is not expressed in any condition, 85% of the genes are highly expressed in minimally one condition, and only 3% of the genes are highly expressed in all conditions. This study comprehensively mapped transcription units (TUs) and grouped 2935 promoters into regulons controlled by various RNA polymerase sigma factors, thus accounting for ~66% of the observed variance in transcriptional activity. Because of the exhaustive variation in expression conditions and the high coverage of the employed tiling arrays, it was anticipated that many new putative regulatory elements would be identified. Indeed, the total number of previously unannotated RNA segments with putative regulatory functions was 1583, compared to approximately 4200 annotated protein-encoding genes. Almost 38% of these RNA segments were asRNAs since they were transcribed from the opposite strand of protein-encoding genes. It was concluded that many asRNAs may originate from spurious promoters or incomplete control over transcriptional termination and do not have a function**.**

In **Chapter 3** the reported regulatory RNA mechanisms were reviewed and this chapter also describes added analyses to the set of putative regulatory RNAs identified in **Chapter 2**. It was proposed that spurious asRNAs can be a source of new post-transcriptional regulation and protein-encoding capacity. From the analysis of evolutionary conservation and predicted

structures of all RNA segments, it followed that the sense-independent segments are the most conserved and structured putative regulatory RNAs. Generally, from reported literature it is apparent that, in contrast to eukaryotes, *B. subtilis* and other bacteria often employ post-transcriptional regulation via the 5' end of mRNA molecules, either by controlling read-though transcription, controlling mRNA degradation, or inhibiting ribosome access. Notably, an exception to the 5' regulatory model seems to be regulation via asRNAs and this is specifically the case for type I toxins-antitoxins, which in *B. subtilis* seem to preferentially overlap at the 3' end (10). A new function of a 5' RNA regulatory element is reported in **Chapter 7**. The work described in this chapter focuses on how a ribosome-mediated transcriptional attenuation mechanism can control the expression of two downstream multidrug ABC transporter genes named *bmrC* and *bmrD* (in short *bmrCD*). Transcriptional read-through into these genes is promoted via formation of an anti-terminator in the 5' regulatory element encoded by *bmrB*. The *bmrB* gene also includes an ORF and the anti-terminator in *bmrB* can only be formed upon slow translation of the BmrB peptide. Since the ribosome is slowed down by the presence of ribosome-targeting antibiotics, such as lincomycin and chloramphenicol, the ribosome provides an intrinsic sensing mechanism for such antibiotics. Upon *bmrCD* induction, these antibiotics could be expelled from the cell by the induced ABC-transporter. This is an example of a clear structure-function relationship of a regulatory RNA molecule, and the results invoke ribosomes as intrinsic sensors of translation speed and regulators of transcription. In this context, it would be interesting to see what the effects of low doses of ribosome-targeting antibiotics would be on the genome-wide translation speed. This could, for instance, be quantified with the novel sequencing-based technique of ribosome profiling (11).

Small regulatory RNAs in B. subtilis

SrRNAs are independently transcribed RNA segments that act on their target mRNAs via short, imperfect complementary base-pairing. The goal of the studies reported in **Chapter 4** was to employ the wealth of expression data presented in **Chapter 2** to facilitate the identification of srRNA functions. To do this, the aspects of regulatory RNA organization listed in **Chapter 3** were applied to a subset of the most likely srRNAs. The approach started with the selection of putative srRNAs and proceeded with extensive target predictions. Subsequently, five analyses were performed with the goal of identifying the most likely srRNA targets. These analyses included an evolutionary conservation analysis of both srRNAs and their predicted targets, target enrichment analyses on the target predictions, two expression correlation analyses computed over a 104-condition expression space, and selection of those srRNA-mRNA pairs that are co-expressed. The presence of an sRNA-target pair in one of these analyses led to the attribution of a so-called flag to this pair. Four flags could be assigned to 43 srRNA-target pairs (Table 2 from Chapter 4). Interestingly, these srRNA-target pairs included RnaC/S1022-*abrB*, and *abrB* was subsequently reported as a true target of RnaC/S1022 in **Chapter 5**. Nevertheless, it is presently still to some extent unclear what the real value of these target flags is. This may become more clear over time when more srRNA-target interactions will have been identified. Such studies could be facilitated by the mining of the predictions presented in this chapter, in a similar way as what was described for FsrA/S512, SR1, S462 and RsaE/S415.

In **Chapter 3** it was noted that the RNA chaperone Hfq does not play the central role in *B. subtilis* as was observed for Hfq in Gram-negative bacteria. It is presently unclear whether there is another RNA chaperone protein central to RNA regulation in *B. subtilis*, or if perhaps there is no requirement for one. Alternatively, designated condition-specific RNA chaperones may have evolved for RNA interactions that do require a chaperone. Notably, the two identified deregulated or direct srRNA-target interactions described in this thesis for the *odhA* and *abrB* genes can now be exploited to identify such general or specific chaperones by screening available

B. subtilis deletion mutant libraries with the respective reporter constructs.

Two aspects of srRNA regulation that are discussed in **Chapters 3** and **4** are sequence and functional conservation. Regulatory RNAs are known to exhibit limited sequence conservation and this was in general also observed in *B. subtilis*. For instance, only one srRNA, RsaE/S415, is conserved to such an extent that it is present in the genomes of both *S. aureus* and *B. subtilis*. All sense RNA segments clustered in three groups of conservation level and this reflects the relatedness of the included genomes. Functional conservation is the conservation of the regulatory *mechanism* rather than the conservation of the regulator's sequence. This is illustrated by the *E. coli* RyhB and *B. subtilis* FrsA/S512 srRNAs, which mediate the iron-sparing response. In both species these srRNAs are Fur-regulated, function in the iron sparing response, and they even have a shared target (*sdhCDAB*) (12, 13), without sharing any sequence conservation. The reason for this might be that srRNA regulation provides such a significant evolutionary advantage that it evolved multiple times. This could relate to issues of response time or other subtle aspects of this regulation. However, it may be more likely that these two srRNAs do share a common ancestor, but that this is not visible anymore at the sequence level after approximately 1 billion years of evolution (14). Sequence conservation is still detectable for the RsaE molecules of *S. aureus* and *B. subtilis*. Possible conserved functions for this srRNA were explored in the studies described in **Chapter 4**, and it was suggested that these include cofactor metabolism, energy transport, and lipid metabolism.

The concept of srRNA regulons was introduced in **Chapter 3** and explored via enrichment of predicted targets with a certain annotation category in **Chapter 4**. The only *B. subtilis* srRNA that has a confirmed regulon is FsrA/S512 with all of its targets related to the iron-sparing response (15). The results presented in **Chapter 4** show that this FsrA/S512 function is also predicted from the employed target predictions combined with functional enrichment analysis. In addition, five other regulon members are suggested based on previously published proteomics data and target predictions. It seems likely that other srRNA regulons are yet to be uncovered in *B. subtilis*, possibly with help from the predictions presented in **Chapter 4**. However, srRNAs are not required to have multi-target regulons, as long as the regulation of a single target provides sufficient evolutionary advantage for the respective organism. It is also possible that srRNAs have multiple targets, but that these are not regulated at the same time. For instance, when target A is expressed in a condition where the respective srRNA is also expressed, while target B is absent, only target A will be regulated by this srRNA. Another possibility for condition-dependent target regulation is the presence of condition- and srRNA-mRNA-specific RNA chaperones. Such chaperones seem to be involved in the regulation of FsrA/S512 targets (16).

The expression levels of an srRNA and its target represent an important aspect of srRNA regulation. However, this aspect is often interpreted too simply. Most focus is on those srRNAs that completely silence their target mRNAs by being present at much higher levels than these. However, according to theoretical studies these srRNA-target pairs are only one out of three categories that should be considered (17, 18). In the second category of srRNA-target pairs, the level of the srRNA is much lower than that of the target. The target is thus almost unaffected by the presence of the srRNA and is basically normally expressed. The third scenario is the intermediate one, where the production rate of the srRNA and the mRNA are comparable and it is this scenario that allows fine-tuning of target expression (18). In the latter case, every small quantitative change in the expression level of the srRNA and the mRNA will lead to quantitative changes in target expression. Therefore, the intermediate scenario is most sensitive to fluctuations and noise (18). Because this deals with noise generation, the latter situation seems applicable to the interaction between RnaC/S1022 and *abrB*, which was reported in **Chapter 5**. In the studies presented in this chapter it was shown that the post-transcriptional RnaC/

S1022-*abrB* interaction allows *B. subtilis* to increase the cell-to-cell variation in AbrB protein levels, despite strong negative autoregulation (NAR) of the *abrB* promoter. The evolutionary constraint on NAR of the *abrB* gene provides a clear explanation why AbrB expression noise is regulated post-transcriptionally. The srRNA-induced diversity in AbrB levels is shown to result in heterogeneity in growth rates during the exponential growth phase. This growth rate heterogeneity is conceivably of physiological relevance since slowly growing bacterial cells are generally less susceptible to antibiotics and other environmental insults than fast growing cells (19, 20). This suggests that the AbrB noise level has been fine-tuned in evolution, possibly as a bet-hedging strategy to deal with environmental insults. **Chapter 5** describes a stochastic simulation model of srRNA regulation, which is an additional example of the fruitful marriage between computational and biological analyses. It shows that a computational model can help unraveling a complex and perhaps not intuitive biological phenomenon, and that it can even give suggestions to dissect different aspects of the phenomenon.

The research described in **Chapter 5** makes RnaC/S1022 the third srRNA for which a direct target has been identified in *B. subtilis*. The *abrB* gene was thus also identified as an evolutionary conserved RnaC/S1022 target. Such evolutionary conserved target predictions were performed for all putative srRNAs as detailed in **Chapter 4**. Focusing on such conserved predicted targets reduced the number of considered RnaC/S1022 targets to 47, from a total of 147 predicted targets based on the *B. subtilis* 168 genome sequence. For all selected putative srRNAs this type of analysis reduced the number of considered targets from 11419 to 1843. Notably, such large numbers of targets are predicted, because srRNA-target interaction regions can be very small. Since 6 to 8 complementary base-pairs can be enough to confer regulation, these binding motifs will occur at random multiple times in the genome. The reduction in the number of possible targets to be considered for RnaC/S1022 made it possible to explain the observed growth phenotype of the RnaC/S1022 mutant via deregulation of a single target (*abrB*) and shows the value of srRNA target predictions if they are used rationally. Studying conserved predicted targets and the presence of multiple of these targets involved in the same process (which leads to functional enrichment) can also help to predict conserved functions for srRNAs as was the case for the suggested role of RsaE/S415 in lipid metabolism as described in **Chapter 4**. Focusing on srRNA-target interactions that are conserved in evolution can thus be helpful in reducing the number of possible srRNA targets.

Another interesting possibility is to use the RNA structure for predicting srRNA regulation. This is based on the observation that accessible regions of the srRNAs and their targets in Gram-negative bacteria are more likely to interact (21). This principle was therefore used to predict the interacting basepair between RnaC/S1022 and *abrB* as presented in **Chapter 5** based on the fact that it is located in the middle of a predicted loop region of the srRNA. In this respect, it should be noted that a remodeling of local secondary structure may take place upon srRNA-mRNA interaction, but this is currently unclear.

Despite filtering methods to reduce the number of considered targets, it will always remain extremely difficult or even impossible to demonstrate that a particular srRNA-target interaction will not take place under any environmental condition. Besides the impossibility to prove the non-existence of an interaction, there may be reasons for srRNAs to interact with targets even without strongly affecting their expression. The possible relevance of these weakly binding srRNA targets was outlined in a theoretical paper by Jost et al. (18). These authors first noted that in both bacteria and mammals, the phenotypes associated with some srRNA deletion mutations are often due to deregulation of one or a few of their targets. This is the case even when the srRNA has additional small effects on (many) other targets. It was subsequently proposed that such weakly binding targets can help buffer fluctuations in srRNA levels in the cell by functioning as a reservoir of the srRNA. This reservoir can help to reduce stochastic

fluctuations in the target level, to make the kinetics of the regulation more robust, but keeping the sensitivity of the regulation. An intriguing prediction following from such considerations is that the *abrB* gene of *B. subtilis*, which was shown to be expressed with large noise due to RnaC/S1022 regulation (**Chapter 5**), could be the only target of RnaC/S1022 in the studied conditions. In future studies, this prediction could for instance be tested by pulsed overexpression and subsequent RNA-seq analysis, followed by a comparison of the differentially expressed srRNAs with extensive target predictions, for instance those described in **Chapter 4**.

Antisense RNAs in B. subtilis

The asRNAs of *B. subtilis* have been defined as those RNAs that overlap with more than 100 bases (or >50% of their length) with protein-encoding genes due to transcription from the opposite DNA strand (22). Four types of asRNA regulation have been distinguished in **Chapter 3**, although the diversity in asRNA regulatory mechanisms seems to be virtually endless. An important consideration for the study of functional asRNAs is that transcription of asRNA does not automatically lead to regulation, because of the structural complexity of RNA molecules. This means that it is possible that even completely complementary sequences can fold in such a way that no effective duplex formation between the two is possible. In addition, and analogous to the discussion on srRNA above, also the conditional-dependency of sense and asRNA expression is of course relevant for effective asRNA regulation to take place. The same holds for the expression level of the asRNA compared to the sense RNA. The three scenarios of expression level that were explained above, asRNA >> mRNA, asRNA ~ mRNA, and asRNA < mRNA (18), are also here expected to lead to different outcomes, including the influence of asRNA regulation on target expression noise. These considerations may help to suggest a physiological relevance for a decrease in sense RNA levels by an asRNA, for instance in the cases of the three full-length asRNAs reported for *B. subtilis* (23, 24, 25). Decreasing the level of transcripts can be advantageous in the conditions in which the asRNA is expressed. Such decreases in sense transcript levels can also increase the influence of fluctuations on the expression of the gene, which can be advantageous for some genes. Another interesting suggestion made by Noone et al. (25) is that creating instability of the mRNA, which could be an asRNA function, makes a transcript more responsive to changes at the transcription level. When a promoter is shut off because of a phenotypical adaptation its transcript level should also be directly reduced, but the speed of this response will depend on the stability of the transcript. Decreasing mRNA stability will thus result in a faster response to the signal. This situation is analogous to the type I antitoxins where the antitoxin asRNA is more instable than the sense mRNA encoding the toxin (10). Therefore, whenever the chromosomal region of the toxin-antitoxin is lost, the toxin mRNA will remain present longer than the antitoxin and this will eventually kill the cell. This is an example of the importance of controlling mRNA stability and the possible roles for regulatory RNAs in doing so. Relevant for studying this type of interactions will be the, hopefully, soon available genome-wide database of RNA half-lifes in *B. subtilis* (Mäder and Nicolas, personal communication), and it will be very exciting to analyze the included data for the possible influences of regulatory RNAs on mRNA half-lives.

Another way of asRNA regulation is not dependent on the sequence of the asRNA, but on its genomic context. An interesting context-dependent mechanism is for instance promoter collision, where the RNA polymerase complex from a convergent asRNA promoter can push away the RNA polymerase complex from the promoter driving transcription of the sense RNA (26). It may be that the asRNA regulation of *rpsD* by S1136-S1134 as described in **Chapter 6** proceeds via such a promoter collision mechanism. However, experimental proof for such a mechanism is difficult to obtain and is now solely founded on the observation that induction of the asRNA in *trans* does not complement the asRNA-dependent regulation observed in *cis*. It

would be interesting but challenging to, for instance, apply atomic force microscopy to visualize such a regulatory mechanism. The question that remains is what the evolutionary rationale for *rpsD* downregulation via an asRNA is. The essential *rpsD* gene is regulated by the housekeeping sigma factor SigA and expressed almost constitutively under all conditions. An additional level of control is RpsD's post-transcriptional negative autoregulation via the structured 5'UTR. This autoregulation allows the cell to prevent synthesis of (many) more RpsD molecules than those taking part in ribosome formation. This control has a consequence for signal integration, an aspect of RNA regulation that was discussed in **Chapter 3**. In case the signal of specific downregulation upon stress is advantageous to the cell, this has to be transduced by a specific factor. Autoregulation of RpsD can tune the number of proteins made from every mRNA (burst size), but not the level of this mRNA. This can, however, be achieved with an asRNA or upstream-binding protein factor. However, upstream binding of a protein factor might be complicated by the autoregulatory 5'UTR structure. It thus seems that asRNA-mediated downregulation of *rpsD* may have been easier to evolve. In turn, the reason for downregulation of the small ribosomal subunit, the consequence of *rpsD* downregulation, is suggested to be reducing the translational capacity to prevent translation of spurious and possibly toxic peptides upon ethanol stress. There are at least two possible origins for these (hypothetical) spurious peptides. The first is that they are made during translation from normal mRNAs by translation errors that are specifically induced by ethanol. This happens via the direct interaction of ethanol with the small ribosomal subunit's decoding region (27). The second possibility is that these peptides arise from - possibly spuriously expressed - asRNAs which are very abundant under conditions of ethanol stress (22). The cause for this increased antisense transcription may be the effect ethanol has in downregulating the *rho* gene, which encodes the transcriptional termination protein Rho (22).

One additional interesting area of study would be to determine whether accurate secondary structure measurements could help to predict base-pairing sense-asRNA interactions. It has been noted in **Chapter 3** that the present-day secondary structure prediction algorithms (like RNAfold which was used in this study), have a strongly reduced prediction certainty for sequences longer than 700 nucleotides (28). Many antisense segments are much longer than 700 nucleotides (**Chapters 2** and **3**) and their predicted secondary structures at the level of individual base pairs are thus unreliable. There are now sequencing-based tools available for the genome-wide measurements of secondary structure (28), which could be applied in *B. subtilis* and integrated with the extensive expression data generated by Nicolas et al. (22). Thus, it may become possible to, for instance, identify kissing-loop interactions between two RNA molecules based on accurately determined secondary structures. However, it may well be the case that RNA-RNA interactions will remain very difficult to model for many more years to come.

General notes and outlook

Some very general goals for research on bacteria were formulated in **Chapter 1** of this thesis. These are the control of bacteria in infectious diseases, preventing or combating the occurrence of antibiotic resistance, understanding bacterial physiology to use bacteria as pesticides or fertilizers, and understanding fundamental biological processes to better engineer bacteria to use them for our benefit. The understanding of one fundamental process, namely that of post-transcriptional regulation of gene expression via RNA, was the topic of this thesis. The last part of this general Discussion section will focus on a (non-exhaustive) list of global aspects for the here called systems understanding and engineering of *B. subtilis* (or other bacteria). Figure 1 summarizes four aspects of such systems understanding in order to facilitate the engineering of bacteria at the systems level. The first of these aspects is to comprehend gene functions (or delete the unknown). *B. subtilis* is one of the best understood microorganisms and arguably the

best-understood Gram-positive bacterium, but approximately half of the around 4200 protein-encoding genes have still no assigned function. As was reported in this thesis, this percentage is much larger for regulatory RNA genes. Many of the essential genes also have no annotated function and finding functions for these is required for a complete understanding of the cell (29). An alternative method for identifying functions of the non-essential genes is to delete the unknown genes. This approach was taken with the goal of refining a large-scale metabolic model (30). Combining single deletion intervals led to the deletion of ~35% of the *B. subtilis* 168 genome (Tanaka & Noirot manuscript in preparation). The resulting collection of deletion strains was further analyzed in a European research project for synthetic biology (BaSynthec) in which our group was involved.

The second aspect for systems understanding and engineering is to control and tune population heterogeneity. Heterogeneity within an isogenic population is important for differentiation and resistance development in many organisms, and *B. subtilis* is often hailed as the least complex model for studies on such processes. Natural competence, motility, protease secretion, and sporulation were known heterogenic phenotypes in *B. subtilis*. The studies presented in **Chapter 5** uncovered that diversity in growth rates is also an actively regulated heterogeneous phenotype, and for the first time implicated srRNAs as molecular players in creating such cell-to-cell diversity. Population heterogeneity is undesired in an industrial setting, because it may lead to only a subpopulation of cells producing the product of interest. It is therefore the industries goal to create homogeneously producing populations and the methods and considerations presented in **Chapter 5** can give clues how to approach this. Especially the observed inconsistency between promoter and protein expression noise measurements are important to keep in mind. It is also not excluded that a heterogeneous culture with a subpopulation of superproducing cells can in total yield more product than a homogeneous population, especially if the non-producing cells provide a beneficial function for the superproducing cells.

A large advantage of using Gram-positive bacteria, such as *B. subtilis*, is that these can directly secrete their products into the extracellular medium. This feature strongly eases the downstream purification of the product. There are multiple secretion systems in *B. subtilis* and understanding their action and regulation might provide leads to make these systems even more efficient, especially in accepting heterologous substrates. Work in our lab focuses on defining the role of the twin-arginine translocation (Tat) secretion system, which is especially interesting since it is able to transport folded proteins over the cell membrane. These Tat-dependent substrates have to fold in the cytoplasm because they have to complex with co-factors or depend on the intracellular milieu for folding. No regulatory RNAs that affect secretion processes have

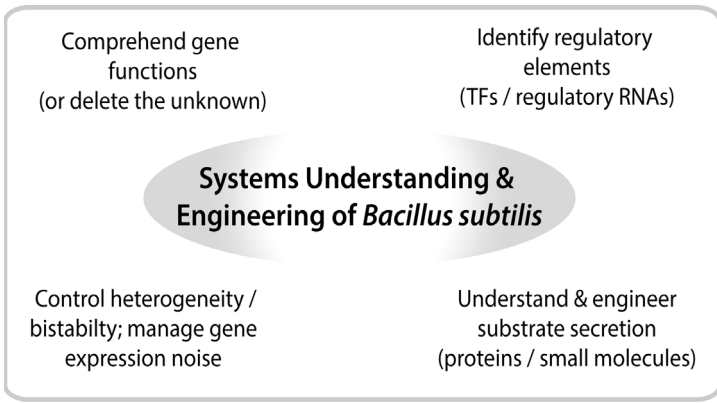

Figure 1. Four required areas of study to enable systems understanding & engineering of *Bacillus subtilis*.

been identified up till now, and this is still an open area of study.

The fourth aspect of systems understanding is to understand and control regulatory elements. These regulatory elements can be protein transcription factors (TFs), which have been quite extensively mapped in *B. subtilis*, or regulatory RNAs. Understanding regulatory elements will facilitate the controlled expression of (heterologous) products, multi-gene pathways, and facilitates control over cellular behavior. A thorough understanding of regulatory elements is thus crucial for systems engineering and synthetic biology. Since many TF regulons are not yet completely identified, deletion libraries of transcription factors can be screened with for instance a focus on substrate expression and secretion. A similar approach could be conceived for srRNAs as well. The research presented in this thesis was specifically aimed at finding the biological function of the regulation by particular regulatory RNA molecules. Since these functions were some of the first to be identified for *B. subtilis*, this research was important to increase our understanding of the roles of regulatory RNAs in this Gram-positive model organism. Once our knowledge of the molecular mechanisms of gene expression regulation via regulatory RNAs has further expanded, it will become possible to design a target gene under control of, for instance, an srRNA. To do this, only the region that is proven to interact with a native target needs to be changed so that it is predicted to interact with the heterologous target. The rest of the molecule can then be kept intact to prevent undesired effects on srRNA stability. Indeed, such an approach was already used for the combinatorial reduction of transcript levels of multiple genes in a metabolic pathway of *E. coli*, and this led to increased product levels (31).

Ultimately, detailed systems knowledge of regulation and sequence-function relationships of regulatory RNAs in model bacteria could help to predict regulatory networks and adaptive responses in newly discovered bacteria. It may then also be possible to predict regulatory functions from expression data or maybe even only from the genome sequence. This would be particularly relevant for studies on bacteria that are impossible to culture. Another perspective is the complete redesign of living systems. One could for instance envision a cell with a synthetic genome, the genes of which are constitutively transcribed at low levels to support low growth rates. Subsequently, one could add in libraries of srRNA molecules to enhance or inhibit translation and screen for enhanced growth rate in an evolutionary engineering strategy. Likewise, it was suggested that the large-scale assembly of metabolite-binding RNA aptamers opens the door to employ these structures as a sort of regulatory RNA 'origami' in cells (32). These are distant but exciting frontiers to push and they will ultimately increase our understanding of the interactions of the molecules that make life possible. Systems technologies are also giving insights into the interesting question whether in living systems everything is perfectly tuned and optimal, or whether cellular components are made in abundance and subsequently degraded when they prove to be not required (11). Thus the question is whether evolution will result in some level of perfection, or whether the end result is only whatever works for being competitive in a certain ecological niche? Because of the illustrated experimental and technological advances, this question can now for the first time be addressed, also in the context of RNA-mediated gene regulation.

References

1. Cech,T.R. and Steitz,J.A. (2014) The noncoding RNA revolution-trashing old rules to forge new ones. *Cell,* **157,** 77-94.
2. RajBhandary,U.L. and Kohrer,C. (2006) Early days of tRNA research: Discovery, function, purification and sequence analysis. *J. Biosci.,* **31,** 439-451.
3. Holley,R.W., Apgar,J., Everett,G.A., Madison,J.T., Masquisee,M., Merrill,S.H., Penswick,J.R. and Zamir,A. (1965) Structure of a ribonucleic acid. *Science,* **147,** 1462-1465.
4. Crick,F.H. (1966) Codon--anticodon pairing: The wobble hypothesis. *J. Mol. Biol.,* **19,** 548-555.
5. Vilfan,I.D., Tsai,Y.C., Clark,T.A., Wegener,J., Dai,Q., Yi,C., Pan,T., Turner,S.W. and Korlach,J. (2013) Analysis of RNA base modification and structural rearrangement by single-molecule real-time detection of reverse transcription. *J. Nanobiotechnology,* **11,** 8-3155-11-8.
6. Sharma,C.M., Papenfort,K., Pernitzsch,S.R., Mollenkopf,H.J., Hinton,J.C. and Vogel,J. (2011) Pervasive post-transcriptional control of genes involved in amino acid metabolism by the *hfq*-dependent GcvB small RNA. *Mol. Microbiol.,* **81,** 1144-1165.
7. Papenfort,K., Podkaminski,D., Hinton,J.C. and Vogel,J. (2012) The ancestral SgrS RNA discriminates horizontally acquired *Salmonella* mRNAs through a single G-U wobble pair. *Proc. Natl. Acad. Sci. U. S. A.,* **109,** E757-64.
8. Prevost,K., Desnoyers,G., Jacques,J.F., Lavoie,F. and Masse,E. (2011) Small RNA-induced mRNA degradation achieved through both translation block and activated cleavage. *Genes Dev.,* **25,** 385-396.
9. Sharma,C.M. and Vogel,J. (2014) Differential RNA-seq: The approach behind and the biological insight gained. *Curr. Opin. Microbiol.,* **19C,** 97-105.
10. Durand,S., Jahn,N., Condon,C. and Brantl,S. (2012) Type I toxin-antitoxin systems in *Bacillus subtilis*. *RNA Biol.,* **9,** 1491-1497.
11. Li,G.W., Burkhardt,D., Gross,C. and Weissman,J.S. (2014) Quantifying absolute protein synthesis rates reveals principles underlying allocation of cellular resources. *Cell,* **157,** 624-635.
12. Masse,E. and Gottesman,S. (2002) A small RNA regulates the expression of genes involved in iron metabolism in *Escherichia coli*. *Proc. Natl. Acad. Sci. U. S. A.,* **99,** 4620-4625.
13. Gaballa,A., Antelmann,H., Aguilar,C., Khakh,S.K., Song,K.B., Smaldone,G.T. and Helmann,J.D. (2008) The *Bacillus subtilis* iron-sparing response is mediated by a Fur-regulated small RNA and three small, basic proteins. *Proc. Natl. Acad. Sci. U. S. A.,* **105,** 11927-11932.
14. Condon,C. and Putzer,H. (2002) The phylogenetic distribution of bacterial ribonucleases. *Nucleic Acids Res.,* **30,** 5339-5346.
15. Smaldone,G.T., Revelles,O., Gaballa,A., Sauer,U., Antelmann,H. and Helmann,J.D. (2012) A global investigation of the *Bacillus subtilis* iron-sparing response identifies major changes in metabolism. *J. Bacteriol.,* **194,** 2594-2605.
16. Smaldone,G.T., Antelmann,H., Gaballa,A. and Helmann,J.D. (2012) The FsrA sRNA and FbpB protein mediate the iron-dependent induction of the *Bacillus subtilis* lutABC iron-sulfur-containing oxidases. *J. Bacteriol.,* **194,** 2586-2593.
17. Levine,E., Zhang,Z., Kuhlman,T. and Hwa,T. (2007) Quantitative characteristics of gene regulation by small RNA. *PLoS Biol.,* **5,** e229.
18. Jost,D., Nowojewski,A. and Levine,E. (2013) Regulating the many to benefit the few: Role of weak small RNA targets. *Biophys. J.,* **104,** 1773-1782.
19. Tuomanen,E., Cozens,R., Tosch,W., Zak,O. and Tomasz,A. (1986) The rate of killing of *Escherichia coli* by ß-lactam antibiotics is strictly proportional to the rate of bacterial growth. *Journal of General Microbiology,* **132,** 1297-1304.
20. Gefen,O. and Balaban,N.Q. (2009) The importance of being persistent: Heterogeneity of bacterial populations under antibiotic stress. *FEMS Microbiol. Rev.,* **33,** 704-717.
21. Peer,A. and Margalit,H. (2011) Accessibility and evolutionary conservation mark bacterial small-RNA target-binding regions. *J. Bacteriol.,* **193,** 1690-1701.
22. Nicolas,P., Mader,U., Dervyn,E., Rochat,T., Leduc,A., Pigeonneau,N., Bidnenko,E., Marchadier,E., Hoebeke,M., Aymerich,S., et al. (2012) Condition-dependent transcriptome reveals high-level regulatory architecture in *Bacillus subtilis*. *Science,* **335,** 1103-1106.
23. Eiamphungporn,W. and 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.
24. Luo,Y. and Helmann,J.D. (2012) A sigmaD-dependent antisense transcript modulates expression of the cyclic-di-AMP hydrolase GdpP in *Bacillus subtilis*. *Microbiology,* **158,** 2732-2741.
25. Noone,D., Salzberg,L.I., Botella,E., Basell,K., Becher,D., Antelmann,H. and Devine,K.M. (2014) A highly unstable transcript makes CwlO D,L-endopeptidase expression responsive to growth conditions in *Bacillus subtilis*. *J. Bacteriol.,* **196,** 237-247.
26. Georg,J. and Hess,W.R. (2011) *Cis*-antisense RNA, another level of gene regulation in bacteria. *Microbiology and Molecular Biology Reviews,* **75,** 286-300.

27. Haft,R.J., Keating,D.H., Schwaegler,T., Schwalbach,M.S., Vinokur,J., Tremaine,M., Peters,J.M., Kotlajich,M.V., Pohlmann,E.L., Ong,I.M., et al. (2014) Correcting direct effects of ethanol on translation and transcription machinery confers ethanol tolerance in bacteria. *Proc. Natl. Acad. Sci. U. S. A.,* **111,** E2576-85.
28. Wan,Y., Kertesz,M., Spitale,R.C., Segal,E. and Chang,H.Y. (2011) Understanding the transcriptome through RNA structure. *Nat. Rev. Genet.,* **12,** 641-655.
29. Commichau,F.M., Pietack,N. and Stulke,J. (2013) Essential genes in *Bacillus subtilis*: A re-evaluation after ten years. *Mol. Biosyst,* **9,** 1068-1075.
30. Tanaka,K., Henry,C.S., Zinner,J.F., Jolivet,E., Cohoon,M.P., Xia,F., Bidnenko,V., Ehrlich,S.D., Stevens,R.L. and Noirot,P. (2013) Building the repertoire of dispensable chromosome regions in *Bacillus subtilis* entails major refinement of cognate large-scale metabolic model. *Nucleic Acids Res.,* **41,** 687-699.
31. Na,D., Yoo,S.M., Chung,H., Park,H., Park,J.H. and Lee,S.Y. (2013) Metabolic engineering of *Escherichia coli* using synthetic small regulatory RNAs. *Nat. Biotechnol.,* **31,** 170-174.
32. Qi,L.S. and Arkin,A.P. (2014) A versatile framework for microbial engineering using synthetic non-coding RNAs. *Nat. Rev. Microbiol.,* **12,** 341-354.