Comparative systems biological analysis of the potential of a high-performance expression platform – *Bacillus pumilus*

Dissertation
for the award of the degree
"Doctor rerum naturalium" (Dr.rer.nat.)
of the Georg-August-Universität Goettingen

within the doctoral program Microbiology and Biochemistry
of the Georg-August University School of Science (GAUSS)

submitted by
Stefani Díaz Valerio

from San José, Costa Rica

Goettingen, 2023

Thesis committee:

PD. Dr. Heiko Liesegang
Department of Genomic and Applied Microbiology, Institute for Microbiology and Genetics

Jun.-Prof. Dr. Jan de Vries
Department of Applied Bioinformatics, Institute for Microbiology and Genetics

Prof. Dr. Jörg Stülke
Department of General Microbiology, Institute for Microbiology and Genetics

Members of the Examination Board:

Reviewer: PD Dr. Heiko Liesegang
Department of Genomic and Applied Microbiology, Institute for Microbiology and Genetics

Second reviewer: Prof. Dr. Jan de Vries
Department of Applied Bioinformatics, Institute for Microbiology and Genetics

Further members of the Examination Board:

Prof. Dr. Jörg Stülke
Department of General Microbiology, Institute for Microbiology and Genetics

Prof. Dr. Rolf Daniel
Department of Genomic and Applied Microbiology, Institute for Microbiology and Genetics

PD Dr. Michael Hoppert
Department of General Microbiology, Institute for Microbiology and Genetics

Prof. Dr. Kai Heimel
Deparment of Microbial Cell Biology, Institute for Microbiology and Genetics

Date of the oral examination: 28.03.2023

"It is not because things are difficult that we do not dare; it is because we do not dare that things are difficult."

Seneca.

Acknowledgements

I want to express my most sincere gratitude to everyone who in one way or another was part of my journey as a doctoral student. The lessons, memories, and experiences that today I carry with me would not have been possible without you.

I am grateful beyond words to Heiko, who started as a supervisor and ended as a true friend of mine. Thank you for the doors you have opened, for your advise, your patience, for believing in me, and for all your support.

To the members of my thesis committee, your time and valuable suggestions helped me navigate this project. To Rolf and all the current and past members of the Department of Genomic and Applied Microbiology who I have the fortune to call colleagues. All of you guided me into learning and becoming a better scientist.

My students, thank you for trusting in me, for all your great and hard work, Anton, Sebastian, Anat, Raphael, Hannah, Robert and Lennart, I am glad for the opportunity to be part of your journey.

To the ABE enzymes team, my gratitude goes also to you. Thank you for hosting me and for allow me to grow ever more fascinated about *Bacillus*. It was a great pleasure to work together and to learn from you. Even during the long fermentations and late-night samplings, I was always supported and accompanied. "To expect the unexpected" is one of the lessons that I take with me.

I am grateful to the awesome members of the IT-team, no bioinformatic work would have been possible without you. Thank you for your expertise, support, and patience with non-expert users.

I found a family outside my home country. Cici, Annette, Nacho, Debbie, Moaz, Blanca, Annabel, Guida, Dani, Jorge, Helmuth, Fiona, Tati, Dirk, Niel, Sofi you kept me sane and helped me see the good side of life even during times of uncertainty. Legion, despite the distance and time, there are things that never change, one of those is true friendship, seeing you again gave me just what I needed before entering the last phase of my thesis.

No sería la persona que soy ahora sin mi familia y su apoyo incondicional. Papá, Mamá, este logro es por ustedes, mis primeros maestros. Cuento en mi vida con grandes mujeres que son una inagotable fuente de inspiración y afecto, abuela Juana, tia Gaby, abuela Telva, gracias!

Finally, to Martin, once again I must say it: a lifetime together will not be enough to thank you for all that you are, and all the love you add to every single day of my life.

Thank you

Contents

List of Figures

List of Tables

List of Abbreviations

asRNA	antisense RNA
ANI	Average Nucleotide Identity
B.	*Bacillus*
bp	basepair
BLAST	Basic Local Alignment Search Tool
BGC	Biosynthetic Gene Cluster
COG	Cluster of Orthologous Groups
DNA	Deoxyribonucleic acid
GO	Gene Ontology
gff	genome file format
GC	guanine+cytosine
HCDC	high cell density culture
h	hours
IS	Insertion Sequence
KEGG	Kyoto Encyclopedia of Genes and Genomes
mRNA	messenger RNA
min	minutes
MGE	Mobile Genetic Element
NGS	Next Generation Sequencing
nt	nucleotide
OD	Optical Density
PCR	Polymerase Chain Reaction
RNA	Ribonucleic acid
rRNA	ribosomal RNA
RBS	Ribosome binding site
RIN	RNA Integrity Number
sRNA	small RNA
TIN	Transcript Integrity Number
TF	Transcription Factor
TPM	Transcripts Per Million
tRNA	transfer RNA
wig	wiggle

Preface

This project was supported by the German Federal Ministry of Education and Research (BMBF) with the grant ID: Fkz.031B057 under the Industrial Biotechnology Innovation Initiative. The particular program corresponds to the Bioeconomy Initiative (UfiB, for "Umsetzungsfördernde Maßnahme in der Initiative Bioökonomie") and aims to investigate and establish more sustainable (waste-avoiding, energy and resource saving) production strategies to replace conventional methods in the manufacturing industry. The framework includes strategic alliances between universities and industrial companies, in this case between Georg-August University, Goettingen and the industrial partner AB Enzymes GmbH.

The present work is part of the base-ground studies required to asses the biotechnological potential of a promising expression platform, *B. pumilus*. This comparative systems biology approach allows to better understand what makes this organism unique regarding other well characterized and established productions hosts, like *B. subtilis* and *B. licheniformis*. Given the close relatedness of these *Bacillus*, it has been suggested that differences in productive performance lay in regulatory elements, such as small non-coding RNAs. Therefore, comparative transcriptomics was proposed as the approach to identify and characterize potential regulators, these findings not only promise to increase the understanding of *Bacillus* biology, but also to aid in the rational optimization of *Bacillus* as industrial production hosts.

The first block of this work illustrates applications of *Bacillus* species in industry and their relevance for current society goals in terms of bioeconomy and sustainability. It explores essential features desirable in any bacterium to undergo further optimization for biotechnology implementations. Not all intrinsic features of *Bacillus* are beneficial in an industrial set up. Sporulation, one of the most studied processes in *Bacillus* species, and one of the reasons for their environmental and evolutionary success is actually problematic. Spores are resilient, and their permanence in industrial environments can lead to contamination, decline of productivity, and higher sterilization costs. On the other hand, genetic accessibility is an essential requirement, not only for expression of recombinant proteins and strain metabolic optimization via genetic engineering, but also to tackle undesirable features, such as sporulation. Genetic accessibility is further explored in this study, with main focus on transformation methods, their optimization, and the barriers to be passed in order to manipulate the genetic material of a strain of interest, like the undomesticated *B. pumilus* DSM27, which could not by tamed during this work. Once a collection of sporulation deficient mutants was established, a set of small-scale fermentations at the facilities of the industrial partner were performed to generate samples for comparative RNA-seq studies.

A second block focuses on comparative genomics approaches to better understand the potential of a bacterium as an industrial productive host. *B. pumilus* genomes were compared, and representatives from *B. pumilus*, *B. subtilis* and *B. licheniformis* were further investigated in a inter-species comparison between emerging and already characterized industrial production workhorses species. From such analysis, putative optimization targets emerged, and more importantly, particular features, which might play a role in *Bacillus* unique adaptation strategies, were highlighted.

Due to recent advancements in sequencing technologies, specially regarding RNA-seq techniques, it has become more and more evident that post-transcriptional regulation also plays a fundamental role in determining a specific organism as a good production host. Consequently, studies based on genomic features are greatly complemented by transcriptomic approaches. A critical step for any successful RNA-seq study is to ensure that the RNA sample is of great quality in terms of purity, yield and integrity. Major efforts were dedicated to investigate how to generate such samples, this was because despite the existence of well established protocols, the methods had to be further tailored for the *Bacillus* samples collected from small-scale fermentations. Eventually, an optimized protocol overcoming common challenges of RNA isolation was generated.

Once good quality RNA samples were obtained and libraries produced, the downstream transcriptomic analysis could begin. For this block of the project, RNA-seq data from *B. pumilus* MS32 was compared with that of *B. licheniformis* MW3. The final section of this project explores state of the art bioinformatic methods for RNA-seq studies. Special attention is given to detection of small regulatory RNAs and expression profiles of *B. pumilus* and *B. licheniformis* during fermentation conditions within small-scale bioreactors. By this approach the bacterial behavior in conditions closer to industrial production is revealed. Moreover, it allows the direct comparative transcriptomic analysis between these two species. Species-specific differences in regulation and transcription of genomic features are discussed. The findings are relevant for selection, development, and optimization of *Bacillus* species as microbial cell factories.

Finally, scientific discoveries must not be restricted to the academic environment, research occurs within an active society context. Therefore, science communication is briefly introduced in Chapter 5 as well as small case of study regarding *Bacillus* for biotechnological applications.

Chapter 1

Introduction

1.1 *Bacillus* species in industrial applications, relevance and requirements

1.1.1 *Bacillus* and its enzymes for a circular bioeconomy

Transition towards a circular bioeconomy is regarded as imperative to address current society goals in terms of environment protection, climate change, and energy supply [349]. While traditional linear economy follows a "take-make-dispose" model regarding utilization of resources and raw materials, circular systems promote re-incorporation, (bio)degradation and re-use of resources, which contributes to a more sustainable development [170, 55]. Fine details about its definition are reviewed, discussed, and criticized elsewhere [173, 63]. Broadly defined, circular bioeconomy can be understood as an intersection, where circular economy principles are applied to bioeconomy models.

Industrial biotechnology plays a key role in the implementation of a circular bioeconomy, as it offers innovations in areas such as energy, fuels, pharmaceuticals, enzymes, and new materials [349, 203]. Within this field, discovery, characterization and development of microbial enzymes for industrial applications is a growing sector due to the increasing demand for sustainable technologies [251]. Enzyme based catalysis is superior to chemical processes due to the high-efficiency rates and selectivity depicted by enzymes, besides their wide range of physical conditions and substrates [251, 315]. Regarded as a green technology, enzymes are now applied in several industries, such as food, feed, textiles, pharmaceutics, and detergents, with new applications being constantly explored.

"Bioeconomy covers all sectors and systems that rely on biological resources (animals, plants, micro-organisms and derived biomass, including organic waste), their functions and principles. It includes and interlinks: land and marine ecosystems and the services they provide; all primary production sectors that use and produce biological resources (agriculture, forestry, fisheries and aquaculture); and all economic and industrial sectors that use biological resources and processes to produce food, feed, bio-based products, energy and services" [55].

Production of industrial enzymes is carried out mainly by microorganisms such as yeast, bacteria, and filamentous fungi as they are broadly available, easy to culture, have a rapid growth rate, and can be genetically engineered to improve desired features. [315]. Among the bacterial hosts for enzyme production, strains of the genus *Bacillus* excel not only at offering a diverse array of enzymes with a wide range of applications, but also at providing some of the most used and well characterized enzyme producers. Particularly, members of the *B. subtilis* clade, such as *B. subtilis* and *B. licheniformis* are widely used in biotechnology [69]. These *Bacillus* are widely distributed in very different environments, reflecting their versatile metabolic capabilities, as well as their capacity to respond and adapt to changing environmental conditions, these features are of interest for industrial applications.

1.1.2 *Bacillus* strains as industrial enzyme production hosts

Implementation of bacterial enzyme producers, like *B. subtilis* and *B. licheniformis*, benefits from some of their natural features which are advantageous within an industrial production set up. For example, in wild environments, such as soil, bacteria from the genus *Bacillus* naturally secrete enzymes to the surroundings. These secreted enzymes facilitate the breakdown of highly polymeric nutrient sources like polysaccharides, nucleid acids, lipids, and proteins, into smaller units. Active uptake systems allow the bacteria to internalize these nutrients to support cell growth. This natural ability to secrete enzymes is further exploited for industrial production. [39, 131].

Secretory capacity: *Bacillus* secretion capacity is superior (up to 20-25g of protein per liter of medium) to that of other organisms and is highly valuable for industrial processes [183]. Secreted products remove the necessity of disrupting the cell in order to obtain the product, making purification simpler and more cost-efficient [197]. In other organisms, such as the Gram-negative bacterium *E. coli* with its limited secretion capacity, the product is accumulated within the cell. This might result in toxic effects for the bacteria, moreover, the intracellular accumulation of product can lead to issues like protein misfolding and generation of insoluble inclusion bodies, both detrimental for production purposes [285].

Biosafety: before implementation of any bacteria as an enzyme production host, it is crucial to ensure that the organism would not represent a safety risk, specially if downstream applications involve food, feed, or pharmaceutical usage. *Bacillus* strains used in industry are checked for exotoxins and endotoxins [67] and have normally the GRAS (Generally Recognized As Safe) status from the U.S. Food and Drug Administration (FDA). Similarly, the European Food Safety Authority (EFSA) recognizes strains from the *B. subtilis* and *B. licheniformis* species as suitable for Qualified Presumption of Safety (QPS) assessments. Examples of qualified applications of *B. subtilis* and *B. licheniformis* strains are given in tables 1.1 and 1.2. For a strain of interest, earning such status requires: establishment of the correspondent identity, no evidence of toxigenic activity, and absence of acquired resistance to relevant antibiotics. Additionally, specific applications of enzymes produced by these organisms are evaluated for safety as well. For example, food enzymes are assessed for systemic toxicity and similarity to known allergens, and the final preparation must be free from viable cells and recombinant DNA. In the case of expression systems based on bacteria like *E. coli*, potential pathogenicity is a limiting factor in food industry applications [355].

TABLE 1.1: Examples of *B. subtilis* industrial applications with QPS (Qualified Presumption of Safety) status by the EFSA (European Food Safety Authority).

Strain	Application	Reference
NBA	production of a-amylase for baking	[101]
ROM	production of maltogenic a-amylase for baking	[98]
DSM 28343	spore preparation as zootechnical additive for calves	[2]
PB6	spore preparation as feed additive for chickens	[11]
TD160	production of endo-1,4-b-xylanase for baking	[100]
NZYM-AK	production of pullulan 6-a-glucanohydrolase for starch processing	[99]
DP-Ezm28	production of endo-1,3(4)-a-glucanase for alcohol production and brewing processes	[102]

TABLE 1.2: Examples of *B. licheniformis* industrial applications with QPS (Qualified Presumption of Safety) status by the EFSA (European Food Safety Authority).

Strain	Application	Reference
NZYM-VR	production of phospholipase C for degumming of fats and oils	[102]
DP-Dzb52	production of α-amylase for starch processing, brewing and distilled alcohol production	[94]
NZYM-BC	production of α-amylase for six food manufacturing processes (starch processing, alcohol distillation, brewing, cereal-based processes, sugar production and juice production)	[95]
NZYM-BT	production of β-galactosidase for milk processing	[96]
DSM 28710	spore preparation as zootechnical additive for turkeys and minor poultry species	[97]

Growth rate: for a cost-efficient industrial production of enzymes, it is advantageous to utilize an organism with a rapid growth rate even on cheap carbon sources, as productivity is linked to substrate consumption. More specifically, volumetric productivity is defined as units of product generated per volume and time, this depends both on cell concentration and the specific productivity of the organism [175]. Industrial *Bacillus* strains successfully grow and secrete enzymes from low-cost substrates, even from agro industry residues such as cassava wastewater, molasses, feather waste, corn steep liquor, and wheat straw [61, 21, 249], which not only saves costs, but also is relevant within a circular bioeconomy framework.

High cell density cultures: it is not only high growth rates, which translates to shorter fermentation times [293], but also the ability to reach and tolerate high cell densities what is desired of a good production host. High microbial biomass often relates to high productivity [308]. High cell density culture (HCDC) entails conditions which might result detrimental for the bacteria, such as nutrient and oxygen limitation, osmotic stress, toxicity caused by accumulation of metabolic byproducts, raising temperature, and mechanical stress due to agitation [306]. Nevertheless, HCDC is still regarded as a prerequisite to maximize productivity. By bioprocess engineering, some of the negative factors might be mitigated, however, even with controlled conditions, not every bacteria is able to thrive to the desired biomass. Members of the *Bacillus* genus are robust by nature, as they evolved to respond and adapt to a wide range of changing and stressing environments, such as soil, where nutrient availability is limited, or a haystack or compost pile, where they can directly colonize the niche and make use of the nutrients. This versatility allows them to grow and become some of the must abundant bacteria within environmental populations. This feature has been further exploited, for example, Voulanto et al, achieved a cell density of 56g/l of *B. subtilis* in a fermentation to produce phytase [332]. More recently in a study to optimize a self-inducing expression system for *B.*

subtilis 73 g/l of dry cell weight were obtained [345].

Supporting background knowledge: next to secretion capacity, safety, growth-rate, and robustness to tolerate fermentative conditions, another desired feature of a bacteria used for industrial production is a sound knowledge base of the organism. *Bacillus subtilis* is the best studied Gram-positive bacteria, with more than half a century of research building insight on its physiology, genetics, and biochemistry. Its complete genome was published already in 1997, and since then it has been complemented by "omics" approaches such as transcriptomics, metabolomics, and proteomics. This research body has been integrated into dedicated public databases such as SubtiWiki [250], and MetaCyc [167], more resources are described elsewhere [125]. Notably, a big portion of this information can be translated to closely related *Bacillus* species and therefore contributes to develop them as additional production hosts. A known example is *B. licheniformis* [330].

Not all natural features of *Bacillus* are beneficial for biotechnological applications, such as enzyme production. The ability to form spores is one of *Bacillus* best adaptations to endure harsh conditions (temperature, desiccation, UV radiation, even extraterrestrial settings) and prevail in the environment [183]. Nevertheless, unlike other advantageous features, which are transferable and desirable for industrial production, sporulation is actually problematic for several industrial applications, because it affects productivity, and might lead to contamination making it hard to sterilize huge fermentation devices. Therefore, sporulation is often engineered out of *Bacillus* strains used as productions hosts.

B. pumilus, like *B. licheniformis*, is another close relative of *B. subtilis*, it belongs to the *B. subtilis clade*. Because it shares many properties of interest with *B. subtilis* and *B. licheniformis*, it has recently caught research interest as a promising industrial production platform [334, 90, 184, 282]. Therefore, this study aimed to further evaluate *B. pumilus* potential as an industrial workhorse. By understanding the endogenous characteristics of a productive strain, their performance and product yield can be optimized, given one more additional condition: genetic accessibility [151].

1.1.3 Prerequisite: genetic accessibility

Genetic accessibility is essential to gain insight into the metabolical and physiological features of any bacterium of industrial interest. Moreover, genetic manipulation techniques, which have been fundamental for microbial biotechnology development, allow to engineer and optimize beneficial traits for industrial production. From this need, several methods for genetic transformation have been developed and improved over the years, sometimes requiring labor-extensive experiments in order to adapt them to a particular strain.

Competence refers to a physiological state in which the cells are able to uptake exogenous DNA, this is physiologically and genetically determined for a specific bacterial strain [78]. In the case of *B. subtilis* the competent state starts to develop at early stationary phase as result of nutrient limitation and quorum sensing signals [209]. Competence of *Bacillus subtilis* has been subject of study since 1960ies, with first reports by Spizizen [302]. Transformation can be understood as the active acquisition and incorporation of extracellular DNA into a host cell, artificial transformation implies a previous procedure (such as electroporation, heat-shock or chemical treatment) when the cells do not exhibit natural competence, for a detailed review on competence and transformation in *B. subtilis* see [209]. The transformation methods relevant for this study are briefly introduced:

Protoplast transformation

Principle: bacterial protoplasts are generated when the cell wall is removed by enzymatic digestion [15]. The peptidoglycan layer is regarded a barrier for DNA uptake. Therefore, by exposing the cell membrane directly to DNA in the presence of polyethylene glycol (PEG), transformation is facilitated. After DNA uptake, a recovery step of incubation in rich media allows the bacteria to restore its cell wall [49, 192]. Protoplast cells are very delicate and extremely sensitive to osmotic and mechanical stress, consequently, for successful transformation careful handling is required.

Conjugation

Principle: the transmission of DNA from a donor to a recipient bacterium by mating is known as conjugation [141, 15]. The donor cell carries a conjugative plasmid encoding elements for expression of a conjugative pilus, synthesis, and transfer of DNA to the recipient cell. One of the first reports for conjugation in *E. coli* traces back to 1946 [189]. Since then, it has been studied for its role in horizontal gene transfer of antibiotic resistance, toxin, and virulence genes in bacteria [222].

Tribos Transformation

Principle: first introduced in 2009, the tribos transformation is based on the recently described Yoshida effect. Briefly, a sliding friction force is applied to a colloidal solution of a nanosized acicular material and bacterial cells over an hydrogel (such as an agar plate), this force allows the acicular material to penetrate the cells. The effect is used to pierce the membrane and deliver the desired DNA [362, 15]. A promising study reported successful transformation of 8 recalcitrant Gram-positive bacteria *B. subtilis*, *B. megaterium*, *Bacillus spp.*, *E. faecalis*, *E. malodoratus*, *E. mundtii*, *L. lactis subsp. lactis* and *L. lactis subsp. cremoris* [270]. The technique has also been used in mammalian cells [47].

Sepiolite: Several acicular materials are suitable for tribos transformation, for example: carbon nanotubes, maghemite, chrysotile, sepiolite, and chitosan. Sepiolite was selected as it is a cheap, abundant, inocuos soft clay, that unlike chrysotile asbestos, is regarded as biocompatible [270]. Moreover, the physical interactions between sepiolite and DNA have been investigated, showing that its silanol groups facilitate DNA adsorption by formation of hydrogen bonds, also presence of multivalent cations allow electrostatic interactions that further favor DNA adsorption [47]. Within this bio nano composite, the structure and function of the DNA is preserved, whether it is chromosomal, plasmid, single or double stranded. Due to the promising physico-chemical properties of the material, sepiolite is currently investigated for more biotechnological applications, for example due to the reversible interaction with DNA it can be use as a low-cost alternative for plasmid purification [264].

This section provided an introduction of the biotechnological relevance of bacteria from the *Bacillus* genus. Particularly, it explored some of the intrinsic features from *Bacillus* organisms which are advantageous within an industrial set up and make them great production platforms. It also presented the current approaches to gain access to the genetic material of an industrial strain of interest. Those were relevant for the selection and implementation of *Bacillus* strains in small-scale bioreactors from which the samples of this project were generated.

B. pumilus MS32 is a novel isolate with no public characterization report. Therefore, there are open questions around this organism: where does MS32 stands in regards to other *B. pumilus* strains? What features encoded by the *B. pumilus* might be relevant for bioprocess applications? Moreover, what can we learn by comparing this species against other well characterized *Bacillus* species currently implemented as enzyme production hosts? Comparative genomic studies offer a path to answer this questions. The next section introduces comparative genomic analysis required for further characterization, evaluation, and comparison of *B. subtilis*, *B. licheniformis* and *B. pumilus*.

1.2 Bioinformatic approaches for comparative genomics

1.2.1 Average Nucleotide Identity

Calculation of the average nucleotide identity (ANI) has become the gold standard for species delimitation in Archaea and Bacteria [12]. Proposed by Goris et al. [118], the method mimics the experimental procedure of DNA-DNA-hybridization (DDH). First, two genomes are fragmented in-silico and homologous regions are identified between each genome. Second, identity is calculated between query fragments and the corresponding homologous regions of the subject genome, and the final ANI is reported as the mean identity of the pairwise comparison of two genome sequences. The boundary cut-off that determines a species is 95-96%, which corresponds to the DDH value of 70% [12].

1.2.2 Orthology prediction

Orthology inference is regarded as one of the most accurate methods to describe differences and similarities in genomic composition between organisms [109]. Broadly, genes derived by speciation events are refer as orthologs, while paralogs designate genes that evolved by duplication events [343], further definitions and subtypes (like orthogroups, co-orthologous, in-paralogs, out-paralogs, xenologs) of these concepts have emerged by studying intricate evolutionary scenarios and are reviewed elsewhere [180]. A major implication derived from orthology relationships, known as the orthology-function conjecture, is the assumption that biologically equivalent functions are carried out in different organisms by orthologous elements, which facilitates extrapolation of functional annotation between genomes[109]. Although there is evidence and exceptions challenging the notion, it is generally accepted that paralogs are more functionally divergent than orthologs and are linked to specific traits that differentiate a given organism.

1.2.3 Gene clusters for biosynthesis of secondary metabolites

Microorganisms produce a wide variety of specialized metabolites, some of these compounds have been characterized and exploited for antimicrobial, anti-cancer, crop protection, food additive, cosmetic, and pharmaceutical applications, and there is still a vast unexplored potential of natural products and applications to discover [341]. A large portion of these natural products are encoded by chromosomally adjacent genes in biosynthetic gene clusters (BGCs) [221]. BGCs can produce several types of chemicals such as polyketides, nonribosomal peptides, alkaloids, and terpenoids [341]. Among the computational tools for genome mining in search for BGCs, antiSMASH [30] is the most widely used and regarded as the gold standard. It is relevant to identify compounds that represent potential safety risks when bacteria is employed in food related biotechnology. Perhaps even more importantly, these specialized metabolites have also been associated to signaling functions impacting physiology and development of *Bacillus* as well as their ecological communities [278, 124, 86, 73].

1.2.4 Resistance determinants

The spread of antibiotic resistance is an issue of global health concern. By identifying and understanding the diverse mechanisms that confer resistance to bacteria we gain insight on their impact on microbial populations, ecology, and health care [219]. Assessment of antimicrobial resistance is required for any bacteria introduced into the food chain. Microorganisms with the potential to cause human infections and/or to transfer antibiotic resistance genes are a risk for food industry applications given the current health crisis caused by the increasing spread of drug-resistant pathogens [45]. The Comprehensive Antibiotic Resistance Database (CARD) is a public resource consisting of a high quality, expert-curated collection of models, data, and algorithms underlying antimicrobial resistance[6].

1.2.5 Prophage prediction

Viruses that infect bacteria are known as bacteriophages, or phages in short. Phages can impact a host genome and phenotype, giving raise to strain diversification and to acquisition of features affecting fitness, such as virulence or antibiotic resistance, moreover, they can alter expression and regulation of bacterial genes [292, 4, 133]. As part of their lysogenic lifestyle, a phage can integrate into the host genome, were it replicates as part of the bacterial DNA while being protected within the cell environment, a phage in this state is called a prophage. Upon a change in environmental conditions, a prophage can become active, excise from the host genome, and induce cell lysis [4]. Phages are tremendously abundant, their genetic diversity is high and their populations are remarkably dynamic, making their identification and characterization in bacterial genomes a challenging task [133, 292]. However, due to the several impacts resulting from phage-host interactions, characterization of phage content in production strains is required and often an optimization point.

1.2.6 Insertion Sequence elements

Among the mobile elements within bacteria, insertion sequence elements (IS) are characterized by being short DNA regions (∼0.7 to ∼2.5 kbp) that encode only the genes required for their own transposition [206, 286]. They are highly abundant within Bacteria, distributed in wide taxa, and display a great diversity of types and copy numbers of different IS carried by a bacterial genome [351, 60]. IS are classified in families according to the encoded transposase, similarity of inverted repeats (IR), conservation of catalytic site, and organization[297]. The activity of IS, their integration and exchange between bacteria can impact genome structure and gene expression. Therefore, IS are relevant to adaptive evolution of their bacterial hosts [286]. Accurate automatic identification and annotation of IS in genomic sequences can be obtained from bioinformatic approaches and contribute to their study [351].

1.2.7 Prediction of CRISPR-Cas Systems

CRISPR-Cas systems are composed by Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) and associated proteins (Cas). They constitute an adaptive defense system against foreign DNA (in some cases RNA) elements (such as virus and plasmids) and are present in most bacteria [210, 8]. CRISPR-Cas systems are of wide interest due to their applications as powerful genome editing tools, with further implications in bioengineering microbial cell factories by allowing effective genetic modification [152, 211].

1.2.8 Identification of Restriction Modification systems

Restriction-Modification (RM) systems constitute a bacterial defense mechanism against foreign DNA. They consist of a restriction enzyme and a DNA methyltransferase [243]. Incoming DNA is recognized as foreign when it lacks the specific chemical signatures of the host cell and is then targeted for degradation by the restriction endonuclease, while the resident DNA is protected by the sequence specific methyltransferase [329, 146]. RM systems can interfere with successful genetic manipulation of a strain of interest (as it will be shown in next chapters). Therefore, identification of RM systems in bacterial genomes is relevant regarding genetic accessibility as a potential optimization point.

1.2.9 Detection of proteolytic enzymes

Peptidases (also known as proteases, proteinases and proteolytic enzymes) are of special interest of study within *Bacillus*. Not only for their production as technical enzymes, but also due to their impact in the overall cell physiology. Proteases fulfill many functions, broadly divided into processing, regulation, and feeding [132]. For example: removal of truncated or misfolded proteins, degradation of proteins with transitory functions, and turnover of ribosomal proteins during stationary phase. Moreover, protease activities impact the production yield of industrial enzymes. Therefore, in many bacteria implemented as production host, extracellular proteases are inactivated to avoid degradation of the desired product [132].

1.2.10 Identification of signal peptides

Signal peptides are short N-terminal sequences in proteins that target them to enter a secretory pathway. Once a protein is at their targeted location, the signal peptide is removed by specific peptidases [137]. In bacteria, the most important mechanism to export proteins out of the cytosol is the general secretory pathway (Sec), which translocates proteins in unfolded state [320]. Another major export pathway is the twin-arginine translocation system (TAt), a remarkable feature of Tat is the transport of substrates in folded (and even oligomeric) states [247]. Research on signal peptides is relevant for industrial biotechnology applications as it impacts several stages of the entire secretory protein production process. Secretion efficiency is often a limit for high yield production of extracellular proteins, particularly regarding secretion of heterologous products in which an non-native signal peptide is fused to the protein of interest and the resulting product is poorly exported [37, 106]. Screening of libraries consisting of a target protein paired with different signal peptides is one approach to identify which pair results in the best secretion efficiency, and to optimize the yield of a desired product [37].

1.2.11 Functional annotation

Moving beyond nucleotide and protein levels of annotation, the next step in order to gain biological significance from a genome of interest is to place the identified features within a bigger context, to link them to the cellular environment and to physiological processes. Functional annotation is perhaps one of the most challenging steps in deciphering the biology of a given organism [304]. With the ever increasing amount of genomic data becoming available, it was evident that a common framework of classification, nomenclature, hierarchy, in other words, systematization of such complex and multi-layer information was required.

One of the first approaches to address the need for such standardized system was the creation of the Gene Ontology resource in 1998 [58]. To this day, with its monthly updates, it constitutes a growing compendium of knowledge around functions encoded by genes and their products [58]. It has been integrated into other resources such as InterPro, an integrative classification of proteins which links signature models from several databases to provide comprehensive characterization of protein sequences [31, 158, 364].

The Kyoto Encyclopedia of Genes and Genomes (KEGG) [163] offers the KEGG Orthology (KO) database of molecular functions organized in groups of functional orthologs. Three databases constitute KEGG: PATHWAY, GENES and LIGAND [163]. The KO database is large collection of protein families (KO families) which is manually curated and used as reference to transfer knowledge from the KEGG databases to specific genes by using identifiers known as K numbers [13].

Another popular approach for functional annotation is provided by the Clusters of Orthologous Genes (COGs) database [311, 111, 310]. In this database, each COG represents a family of co-orthologous genes, which are determined by an all versus all sequence comparison of the proteins encoded in complete genomes. COGs are constructed based on the notion that if a given group of proteins from distant genomes are more similar to each other than to the rest of the proteins in the corresponding genomes, it is likely that they belong to an orthologous group [311]. These groups are manually curated in case by case basis, which includes further refining in case of multidomain proteins, and close analysis based on phylogenetic trees and visual alignment inspection. Functional categories according to COGs cellular functions are assigned to each group, this feature is useful for whole genome characterization and is recommended by the Genome Standards Consortium [111].

By setting up a comprehensive genomic comparison of *B. pumilus* against other well described *Bacillus*, we can discover species-specific characteristics behind their different behavior and productive performance. However, despite the abundant knowledge that can be generated from such approaches, there is a limitation when the goal is to understand active physiological and metabolic processes that define bacterial behavior. This information is not directly accessible by studying only at the DNA sequence level [353, 347, 129]. Therefore, in order to give a more comprehensive characterization and evaluation of the productive potential of *B. pumilus*, the comparative genomic study was complemented with RNA-seq analysis. The next section introduces transcriptomics and the insights attainable by this approach, with particular focus on RNA-based regulation.

1.3 RNA-seq Analysis

1.3.1 Transcriptomics relevance

Several "omics" approaches have been developed to synergize with genomic investigations, for example transcriptomics, proteomics, and metabolomics methods. Study of bacterial transcriptomes was lagging behind those of eukaryotic ones, mostly due to technical challenges regarding selection and enrichment of mRNA [301]. For eukaryotic organisms polyadenylated mRNAs facilitate its selection by use of oligo dT primers. This is not the case for bacteria, as prokaryotic mRNAs lack of polyA tails [327, 255] and tend to have a shorter lifespan than the eukaryotic counterparts [26, 214]. Moreover, rRNA usually represents more than 85% of the total isolated RNA, which hinders detection of other RNAs molecules with sufficient coverage. Recent developments in rRNA depletion methods allowed to enrich the RNAs of interest [327, 255]. With such challenges overcame, a new era of bacterial transcriptomic analysis begun, also facilitated by innovations in sequencing technologies, which lead to the emergence of sophisticated RNA-seq based studies.

For RNA-seq investigation of bacterial transcriptomes, total RNA is extracted, the rRNA is depleted and reverse transcription is used to generate cDNA libraries, which are then prepared for sequencing. The resulting reads are mapped against the reference genome and coverage is calculated to infer transcriptional activity. [301]. The relevance of high quality and integrity of the RNA sample is discussed in the manuscript "RNA of high yield, integrity and purity from industrial *Bacillus*, an improved method" (Chapter 6). By studying which genomic features are actively transcribed at a given condition, functional understanding of genomic elements and its regulatory networks is gained, this also contributes to optimization approaches for microbial cell factories [347, 353].

RNA-seq studies helped to elucidate the previously unexpected complexity of prokaryotic transcriptomes. It allows single nucleotide resolution and offers a dynamic global view of gene expression on the overall cellular context [143]. Here are some examples of areas in which this method has impacted our understanding of RNA biology in bacteria: (for the detailed reviews see [301, 143]).

- Improvement of genomic annotation by identification of small ORFs, non-coding RNAs and UTRs. This complements, confirms, and expands on gene prediction algorithms. Small peptides are often overlooked by annotation software and represent another source of potential regulators [323].

- Identification of untranslated regulatory regions such as riboswitches, 3′-UTRs, 5′-UTRs, TSS and promoter motifs affecting gene expression.

- Detection of operon structures, similarly with gene annotation, there is a bioinformatic challenge to accurately predict bacterial operon structures, transcriptomic analysis facilitate the identification of transcriptional units [62].

- Discovery of widespread antisense transcription, beyond antisense regulation, the observation of pervasive transcription is now a focus of discussion in which further global functions for RNA processing and DNA repair are being proposed [114, 333].

- Functional plasticity of RNAs, under different conditions functional RNA elements have shown different roles, for example some riboswitches also act as small regulatory RNAs. In *B. subtilis* the RNA RosA functions as a sponge of other sRNAs with different outcomes depending on the interacting partner, it sequesters FsrA, and on the other hand, targets RoxS for degradation [80].

- RNA modifications, specialized RNA-seq techniques allow to investigate and profile base modifications in bacterial RNAs. The emerging field, so called "epitranscriptomics", is expected to grow and develop novel approaches to characterize and understand the biological relevance of such modifications [26, 214].

- Characterization of RNA decay and processing. RNA turnover is necessary to balance transcript synthesis and protein output, also to quickly remove unwanted transcripts and adapt to environmental changes. By RNA-seq approaches, RNAse cleavage, RNA lifetime, transcription elongation, and decay rate can be investigated [143].

- Metatranscriptomics, by this approach we can now uncover what are active elements interacting within a whole microbial community, this is particularly relevant to understand organisms which can not be cultured in laboratory conditions and have eluded characterization.

- Non-coding RNAs, whole-transcriptomic analysis revealed small non-coding RNAs (sRNAs) as major players in post-transcriptional regulation involved in many biological processes, these molecules will be further discussed in this chapter.

1.3.2 Small regulatory RNAs (sRNAs)

Bacterial responses to (internal and external) environmental signals involve coordination and fine-tuning of gene expression profiles. Transcriptional, post- transcriptional and translational networks interplay to create regulatory circuits that mediate the adaptive response. In recent decades, the role of RNA molecules within these networks has been unveiled. Nowadays RNA functions beyond rRNA, tRNA and mRNA are recognized. There is now a rich diversity of modulating functions and complexity in mechanisms by which RNA molecules contribute to gene expression control, impacting several cellular processes such as virulence, stress response, sensing population density, modulation of cell surface composition, and metabolism. This occurs in a manner that resembles the microRNA regulation in eukaryotes [241, 143, 115, 248, 333].

Small RNAs (sRNAs) are now known as the main class of post-transcriptional regulators in bacteria [248, 35]. They are abundantly found in all the major branches of the bacterial tree and in several archeal species [333]. Characterized by a size range of 50-500 nucleotides [324] and being frequently (but not always) non coding [115], they play a regulatory role not only by interacting with mRNAs, but also with targets such as other sRNAs, tRNA precursors, or proteins [115, 35, 80]. sRNAs can originate from a wide variety of biogenesis pathways and depict heterogeneity in structure [333]. The average amount of sRNAs encoded by a bacterial genome is estimated to be between 200 and 300 [323].

In most cases, post-transcriptional regulation of gene expression by small non coding RNAs occurs by complementary base pairing with the target mRNA [241]. This interaction can result in activation or inhibition of gene expression, and for several Gram-negative bacteria requires mediation by a general RNA chaperone (e.g Hfq, ProQ), which promotes RNA-RNA anealing, stabilizes, and unfolds RNA [333]. In Gram-positives, the base pairing seems to occur without the need for a general chaperon or is assisted by other RNA-binding proteins in a case specific manner [35]. In *B. subtilis* there is an Hfq ortholog, however it is not essential for the sRNA-mRNA interactions characterized so far [215, 35]. In a *B. subtilis* strain with a deleted *hfq*, only a moderated effect in transcriptomic activity was observed in comparison with the wild-type strain, Hfq seemed to affect toxin and antitoxin transcripts, but no major impact in central post-transcriptional regulation was reported [127].

Only a few sRNAs are constitutively expressed, transcriptional activity is rather induced under particular environmental conditions [333]. According to general classification, the main categories of regulatory small RNAs are *cis*- and *trans*- encoded sRNAs:

Cis- encoded sRNAs

Also known as *bona fide* antisense RNAs (asRNA), which are located directly in the opposite strand of their target protein-coding gene and therefore, exhibit complete complementary with its target [35]. A well characterized example is the antisense sRNA AprAs. Initially identified in *B. licheniformis*, AprAs is encoded opposite to the protease Apr, a member of the subtilisin Carlsberg family, and negatively regulates its production. When transcription of AprAs was prevented, a four-fold increased in protease expression was reported, this observation is of biotechnological relevance for optimization of industrial production platforms [139]. Other examples of *cis-* encoded sRNAs come from toxin-antitoxin systems in *B. subtilis* [323].

Trans-encoded sRNAs

Trans-encoded sRNAs are structured and translated from a different chromosomal location than its (often multiple) targets. In this case, binding of a seed region, of 6 to 8 nucleotides, but often 10 to 12 nucleotides [333], triggers imperfect discontinuous base pairing between the interacting partners. This limited complementarity gives trans-encoded sRNAs the flexibility to interact with multiple targets using different seed regions. Despite the short and interrupted pairing, there is high specificity reported for these interactions [333]. For example, SR1, the first trans-encoded sRNA identified in *B. subtilis* shares seven complementary regions with *ahrC* mRNA. Upon binding, structural changes around the RBS (ribosome binding site) of *ahrC* mRNA inhibit translation initiation by preventing binding of the ribosomal 30S subunit [136, 323]. AhrC is a transcriptional regulator of arginine metabolism. A second target of SR1 is *kinA* mRNA, sharing also seven complementary regions, this target encodes the main histidine kinase participating in the sporulation phosphorelay system and its regulation is achieved by translation inhibition[322]. Moreover, SR1 also encodes a small peptide, called SR1P, which is involved in modulating RNA degradation [116, 323]. sRNAs that additionally encode small functional peptides are also themed dual-function sRNAs [116].

There are several mechanisms, beyond those previously mentioned, by which the sRNA exerts its regulatory role upon base pairing with its target mRNA, interaction could result in: structural changes around the RBS (ribosome binding site) that result in blocking or exposing it, as well as in stabilization, processing, or targeting the mRNA for degradation, and additionally in target trapping, or promoting premature transcription termination [333, 35, 80, 241].

Moreover, other sRNAs can bind to proteins, for example post-transcriptional regulators, and modulate its action based on protein sequestration mechanisms [333]. For example in *Pseudomonas putida* two sRNAs mimic the mRNA targets of the transcriptional regulator Crc, which mediates catabolite repression. Binding of the sRNAs titrates out the regulator from its targets and leads to functional inactivation [217]. Another well characterized example is the ubiquitous 6S RNA, it is around 200 nt in length and interacts with the RNA polymerase to globally regulate its transcription activity, facilitating the shift from vegetative to stationary phase promoters [51, 323, 48].

Integration of regulatory layers

There is a regulatory interplay between transcription factors (TF) and sRNAs. TFs exhibit transcriptional control over sRNAs genes, and in turn sRNAs can modulate TFs post-transcriptionally [333]. These regulatory units can share targets and be part of the same cellular network shaping the bacterial regulatory landscape. There are differences in how and when these regulators are activated to exert its function [294]. One advantage of sRNAs over protein regulators, is that they act quickly and represent a relatively lower metabolic expense [82]. For example, a study indicated that negative regulation, detected by change in target protein levels, was achieved faster by sRNAs than by TFs [294]. Another difference between sRNA-based regulation and that mediated by TFs, is that the transcription rate of the target does have an effect on the fold activity exerted by the sRNA. Additionally, sRNAs allow a faster recovery once the stimuli is removed [333, 294]. Moreover, these regulators can complement each other, for example when some transcripts are produced despite transcriptional repression, translation can be obstructed by action of sRNAs, achieving complete inhibition of gene expression [294]. Regulation by sRNAs seems more advantageous when fast responses in a short time interval are needed [294]. Bacteria might have developed and interlaced these regulatory layers to be used according to different requirements [333].

1.3.3 Identification of sRNAs and their targets

Transcriptomic investigations have established sRNAs as major regulatory elements in bacteria. However, identification and functional characterization of sRNAs and determination of interacting partner networks remain as open challenges addressed by several developing methods. Initial searches for *trans*-acting sRNAs begun with comparative genomics between closely related genomes by scanning intergenic regions for conserved sequences or indications of orphan promoters and terminator sequences [119, 201, 193]. It has been shown that mosts sRNAs are often restricted to a single organism or within closely related species, therefore identification based on sequence homology, such as BLAST [44], is limited. Moreover, by this approach only evolutionary-conserved sRNAs are at reach, and the analysis depends, of course, on the availability of closely related genomes[193].

The sequence flexibility characteristic of sRNAs is not always reflected by sequence conservation, functional sRNA homologues often demonstrate little sequence similarity, and different sRNAs often present different secondary structures [193, 201]. Given the heterogeneity, short size, and little sequence conservation of sRNAs, computational approaches often take into account secondary structure information [201]. For example Infernal [238] uses probabilistic covariance models to identify members of a RNA family based on sequence and secondary structure information. These covariance models are collected in the RFAM database [159]. Another layer of classification within RFAM are Clans. RFAM Clans represent groups of families which fulfill either of these conditions: 1-) members of the family have a clear common ancestor, but their sequences are too divergent to produce a reasonable alignment; or 2-) the members could be aligned, but are kept as separate families because they are functionally distinct. For example, the Clan CL00112 represents 5 families describing archeal, bacterial, and eukaryotic large ribosomal subunit RNAs [112].

Identification of targets for a given sRNA is regarded as a critical bottleneck for functional characterization [18]. Experimental approaches to study sRNAs and their interaction networks include MAPS (MS2-affinity purification and sequencing), GRIL-seq (Global sRNA target identification by ligation and sequencing), RIL-seq (RNA interaction by ligation and sequencing), RNase E-CLASH (RNase E cross-linking and sequencing of hybrids), CLIP-seq (UV cross-linking and immunoprecipitation), PARIS (Psoralen analysis of RNA interactions and structures) and other RNA-seq specialized techniques (reviewed in [348]). In several Gram-negative bacteria, pulldown of sRNA-mRNA pairs together with the Hfq chaperone facilitates the discovery of interacting partners. Since a general chaperone is not essential for these interactions in Gram-positives, relatively less is known about their sRNAs interaction networks [333]. However, development of alternative techniques (without the need of a protein bait) promises to close this gap. An additional challenge is presented due to the fact that some sRNAs are expressed only in very specific conditions, therefore studies under different conditions might not be able to detect all of the sRNAs encoded in a genome [119].

Computational methods to identify mRNA targets of sRNAs look for complementary seed sequences and evaluate the structure accessibility to determine a minimal hybridization energy for the interaction to occur [348]. Computational prediction of interacting partners is helpful to prioritize and reduce the list of putative targets for experimental verification [41]. The program IntaRNA [41] considers accessibility of target sites and allows user defined seed regions [41]. A limitation of computational analysis is that possible sRNA-mRNA pairs are predicted using parameters reflecting known interactions, therefore novel or yet uncharacterized interactions remain elusive [348].

Regarding software for detection of sRNAs from RNA-seq data, a recent study found that APERO [191], TLA from RNA-eXpress [103], and ANNOgesic [363], gave the best performance compared to other available tools [191]. TLA is not specific for bacterial data, and APERO is better suited for non-fragmented and size selected libraries. Annogesic [363], therefore, represents a suitable option for the analysis of this project RNA-seq data, as it has been successfully employed in several RNA-seq based studies [229, 81, 346, 280, 187]. Moreover it is robust, well documented, and compatible with other tools for RNA-seq analysis, such as READemption [104].

Currently, the best approach consists of a combination of bioinformatic analysis together with experimental validation [193]. For this study whole transcriptomic data was complemented with computational analysis to identify regulatory sRNAs and their putative targets. Evaluation of coverage allows to identify sRNAs based on real transcriptional activity. Target prediction is achieved by implementation of RNAup, RNAplex and IntaRNA within the Annogesic software [363]. Moreover, by studying the transcriptional profile of a given sRNA and that of its putative target, indications of negative or positive regulation can be obtained.

1.3.4 The gap, *B. pumilus* transcriptome

As previously mentioned, *B. pumilus* shares many desirable features with established productive hosts such as *B. subtilis* and *B. licheniformis*, which points to *B. pumilus* strains as an attractive source for novel industrial workhorses. This potential is evidenced by the growing research interest around this organism [184, 344, 313, 282, 90, 334, 199, 66, 149, 108].

By 2021 there were 108 putative *trans*-encoded sRNAs reported for *B. subtilis* [35]. Only a handful of studies have implement RNA sequencing in the characterization of *B. pumilus* [353, 129, 196, 199]. For example, a transcriptomic profiling approach was used to study *B. pumilus* BA06, a promising producer of extracellular proteases, and its metabolic changes at different growth phases [129]. Later on, *B. pumilus* SCU11, a derivative of BA06, was investigated regarding small regulatory RNAs [353]. Those studies, were carried on at laboratory scale, which does not reflect accurately the conditions and challenges that bacteria face within a bioreactor during a productive fermentation. Moreover, deeper taxonomic characterizations found BA06 actually closer to *B. altitudinis* (a closely related member of the *B. pumilus* group) and proposed the corresponding re assignation of the strain [85, 198]. Another study characterized *B. pumilus* Jo2 by microarray-based analysis, however, neither the strain or its genomic sequence are available for further investigations [130].

Consequently, a RNA-seq based characterization on taxonomically accurate, and publicly available *B. pumilus* is still missing from literature. The RNA-seq study proposed by this work contributes to the understanding and characterization of the species *B. pumilus*. Further insight is gained by the comparative transcriptomic approach, by which *B. pumilus* response to the fermentation can be directly contrasted to that of *B. licheniformis*. This highlights inter-species differences in adaptation strategies, regulatory responses, and potential optimization targets. Moreover, since the experiments were carried on at small-bioreactor scale, the resulting data is also useful to gain insight regarding bioprocess optimization for biotechnological applications of these *Bacillus*.

Chapter 2

Materials and Methods

The initial bacterial strains selected for this study were: 1) *B. subtilis* 168, the laboratory type strain which is used as model organism for many studies regarding *Bacillus*. This represents a baseline from which the knowledge base can be related to the other closely related species. 2) *B. licheniformis* DSM13 (type strain, for comparative genomics) and its derivative MW3 (for transcriptomic analysis) [339, 263], a strain engineered for better genetic accessibility and unable to produce viable spores, this species is another promising candidate for industrial production of enzymes. 3) The type strain *B. pumilus* DSM27. Other strains, media, and detailed conditions for the subsequent experiments are found in the Appendix A.

2.1 *Bacillus* germination deficient mutants

Given the disadvantages associated with the presence of spores in an industrial environment, particularly regarding potential contamination and higher sterilization costs, the first step for the comparative transcriptomic experiments here proposed, was to create a collection of *B. subtilis*, *B. licheniformis* and *B. pumilus* strains unable to produce viable spores. This was a prerequisite for conducting small-scale fermentations at the Research and Development facilities of this project industrial partner. As it will be shown in further sections, *B. pumilus* DSM27 was replaced by the novel strain MS32, a soil isolate.

2.1.1 Deletion cassette for the *yqfD* gene

In 2003, researchers disrupted several *B. subtilis* genes, and found that disruptions in the *yqfD* gene produced spores blocked at a late stage of maturation and therefore failed to germinate [91]. Similarly, the homologous gene was found in *B.licheniformis* and its deletion lead to the germination deficient MW3 mutant [263, 339].

From the Genomic and Applied Microbiology Department strain collection, *B. licheniformis* MW3 [339] was obtained. Equivalent mutants for *B. subtilis* 168 and *B. pumilus* DSM27 were not in the collection, and therefore had to be generated as described in the following sections.

SOEing PCR to generate the deletion cassette

SOEing PCR allows to create a recombinant fragment of DNA without the need of restriction sites, ligases, or in vitro synthesis. The method was first described by Horton et al [144]. Briefly, an hypothetical PCR product "A" can be fused to another product "B" by having each of them amplified with a standard primer and a "hybrid" primer. The hybrid primer for "A" contains a 5′overhang not matching the template, but actually complementary to one end of "B". Similarly one primer for "B" is complementary to "A". Therefore, both "A" and "B" amplified fragments are complementary to each other at one end. When the two products are combined in a new PCR reaction, they can denature and re-anneal by the common region, this overlap is extended by the polymerase and the recombinant product is generated.

The deletion cassette for the *yqfD* gene of *B. subtilis* was generated as follows. The flanking regions of the *yqfD* gene (around 1000bp) of *B. subtilis* were amplified with primers containing an overhang matching the ends of the *ermD* gene of *B. licheniformis* (Primers FlankA: SDV0006/ SDV0007, FlankB: SDV0008/ SDV0009). Similarly, *ermD* was amplified with primers overlapping the flank regions (Primers: SDV0004/ SDV0005). These three PCR products were purified (MagSi-NGS $^{\text{PREP}}$ Plus) and mixed together in a 2:6:2 ng/µL ratio of flankA:ermD:flankB and used for the SOEing reaction.

The SOEing PCR reaction was done in two steps, first 15 cycles were run without primers, this allows the overlapping regions to self-prime and produce templates of the desired product. In a second phase, reaction mix containing primers starting closely to the cassette extremes (Primers: SDV0014/SDV0015) were added, and the reaction runs for 25 more cycles. The product was run on an agarose gel and the band of expected size (3183bp) was excised and purified (QIAquick Gel Extraction Kit) so it could be used for transformation experiments.

2.1.2 Germination deficient *Bacillus subtilis* 168

In order to produce a germination negative *B. subtilis* 168 mutant, the *yqfD* gene was replaced by the *ermD* gene of *B. licheniformis* 9945A. The *ermD* gene is a macrolide-lincosamide-streptogramin B (MLS) resistance determinant conferring resistance to antibiotics such as erythromycin. The deletion cassette was generated by SOEing (Synthesis by Overlap Extension) PCR [326].

Transformation

B. subtilis 168 was inoculated in 4 ml of NB medium and incubated overnight (37°C, 180 rpm). The overnight culture was diluted to an OD600 of 0.1 - 0.2 and further incubated until a OD600 of 1 - 1.3. The culture was divided into aliqouts of 400 µL. Different concentrations of the deletion cassette for the *yqfD* gene (300, 500 and 750 ng) were added to each aliquot. After 1 hour incubation, 100 µL of expression mix were added to each reaction and further incubated for another hour. Finally the cells were platted on NB agar plates with erythromycin 5 µg/ml.

Transformants growing on the antibiotic plate were selected for 3 colony PCR reactions, reaction 1 targeted *ermD* (Primers SDV004/ SDV005) with purified cassette as positive control, reactions 2 and 3 controlled for incorporation in the correct position by having one primer within *ermD* and the other upstream or downstream the flanking regions (Primers SDV0010/ SDV0011 and SDV0012/SDV0013), as no positive control was available, genomic DNA of *B. subtilis* 168 was used as template to confirm if unspecific products could be expected, water was used as negative control for every set. The PCR products from 2 successful clones were purified (MagSi-NGS $^{\text{PREP}}$ Plus) and further confirmed by Sanger sequencing.

2.1.3 Germination deficient *B. pumilus* DSM27

Engineering of a germination negative mutant of the type strain *B. pumilus* DSM27 was previously attempted at the Genomic and Applied Microbiology Department by Dr. Sonja Volland and MSc. Katrina Funkner by electroporation and transduction methods without success, already indicating that this strain could be more recalcitrant to transformation than the *B. licheniformis* and *B. subtilis* counterparts. The alternative techniques of conjugation, tribos and protoplast transformations were implemented. A deletion cassette for the *yqfD* of DSM27 was generated in the same way as the cassette for *B. subtilis* 168.

Protoplast transformation

Protoplast transformation of *B. pumilus* DSM27 was done based on the methods described in [49] with the modifications made to optimize the protocol for *B. licheniformis* [339].

Preparation of protoplasts: 25 ml of #416 media supplemented with 5 ml of glycerol based recovery media (GRM) were inoculated and incubated overnight (37°C, 180 rpm) in a 250 ml shake flask. The next day, 35 ml of #416 media supplemented with 5 ml of GRM was inoculated from the overnight culture to an OD600 of 0.25-0.3 and incubated (37°C, 180 rpm) until OD600 reached 0.85-0.9. Cells were collected by centrifugation (4°C, 4000 rpm, 15 min) in a 50 ml Falcon tube and resuspended in 5 ml of SMMP (pre-cooled), this step was repeated one more time and then the washed cells were transferred to a 100 ml shake flask where the lyzozyme solution was added. The cells were incubated in presence of lyzozyme with gently shaking (37°C, 80 rpm) until 85-90% of the cells became protoplasts as determined by light microscopy (maximum 2 h). The protoplasts were carefully transferred to a Falcon tube containing 12 ml of SMMP and gently mixed. Protoplast were harvested by centrifugation (420xg, 12 min, room temperature) and the supernatant was carefully discarded. The protoplasts were gently resuspendend in 3 ml of SMMP and separated in 500 µL aliquots to be used immediately for transformation (for later use, 500 µL of 50% glycerol were added as cryo-protectant).

Transformation: 25 µL of 2X SMM and 25 µL of deletion cassette were mixed in a 15 ml Falcon tube, then the freshly made protoplasts were transferred to this tube. Immediately 1.6 ml of PEG solution were added and the protoplasts were gently mixed at room temperature for 2 minutes. Then, 5 ml of SMMP supplemented with 2% BSA (sterilized by filtration) were added and the sample was centrifuged (8°C, 420 xg, 8 min) and resuspended in 1 ml of SMMP with 2% BSA. Finally, the protoplasts were incubated (30°C, 100 rpm, 135 min) and carefully plated on pre-warmed DM3 agar plates supplemented with erythromycin (5 µg/ml) and on DM3 plates without antibiotic.

Cell wall recovery was checked by light microscopy. The clones were isolated and used for two colony PCR confirmation reactions (Primers: H1_forward/ ControlF and H1_forward/ H1_reverse). The first reaction served as control for incorporation at the expected location and the second to confirm the presence of *ermD*. Additionally, primers targeting the *yqfD* gene were created to attest that the gene was effectively deleted. Several transformation rounds were attempted with varying concentrations of DNA (100, 200, 300 and 500 ng/ul), with longer recovery times (up to 290 minutes) and by reducing agitation.

Conjugation

Transformation by means of conjugation for *B. pumilus* DSM27 was done based on previous studies employing the method to transform *B. licheniformis* and related bacilli [263, 138].*E. coli* S17-1 pV2 was used as donor, the strain carries an empty conjugation vector conferring resistance to kanamycin and can be used as shuttle vector for genetic engineering of wild *Bacillus* strains [138]. *B. subtilis* Δ6 was used as control of the method.

Conjugation: the first step was inoculation of 5 ml of LB broth with DSM27 and *B. subtilis* Δ6, and *E. coli* S17-1 pV2, the media for the donor strain was supplemented with kanamycin 25µg/ml, the cultures were grown overnight (37°C, 180 rpm). Fresh LB broth, supplemented with kanamycin for the donor strain, was inoculated from the overnight culture to an OD600 of 0.1, and further incubated until OD600 reached 1. The donor cells (2 ml) were harvested by centrifugation (6000g, 1 min) and the supernatant discarded, afterwards, 2 ml of recipient strain culture was centrifuged on top of the previous pellet. The resulting mixed pellet was resuspended in 200 µL of LB medium and spread over LB agar plates without antibiotic, then left at room temperature until dry. The plates were incubated overnight at 30 °C. The next day cells were collected by washing the plates with LB media and transferred to new LB agar plates containing 25µg/ml kanamycin and 20µg/ml polymyxin and incubated overnight at 30°C.

Tribos Transformation

Based on previous studies utilizing the tribos method for different organisms [270, 269, 362], the following protocol was devised for *B. pumilus* DSM27.

Transformation: 20 ml of LB broth were inoculated and incubated overnight (30 °C, 180 rpm). The next day 40 ml of fresh LB broth in a 500 ml shaker were inoculated from the overnight culture to an OD600 of 0,3 and incubated (30 °C, 180 rpm) until OD600 doubled. The cells were harvested by centrifugation (10 min, 3000g, room temperature) and resuspended in 700 µL of NB, this volume was divided into aliquots of 100 µL and pelleted again. Each pellet was resuspended in 80 µL of sepiolite-DNA suspension (DNA, 0.01 % sepiolite, 100 mM CaCl2) and incubated at 55 °C for 2 min. Finally, the cells were agitated by vortexing for 1 minute before transferring to a dry LB agar plate supplemented with erythromycin (5 µg/ml). In order to elicit the Yoshida effect, the plate was placed on a magnetic stirrer and a sterile stir bar was spun on top of the agar for 1 minute.

The method was tested with 300 ng or 500 ng of DNA, either deletion cassette or plasmids (pRH18, pRH21 and pMR13, kindly provided by Dr. Robert Hertel). Another test combined 500 ng of cassette with 50 ng of genomic DNA extracted from DSM27. Cell pellet size and incubation parameters were also varied in an attempt to optimize the protocol for DSM27.

2.2 Bioinformatic for comparative genomics analysis

Figure 2.1 summarizes the steps for comparative genomics analysis performed on the *Bacillus pumilus* MS32, *B. subtilis* 168 and *B. licheniformis* DSM13 genomes and other *B. pumilus* genomes available at NCBI. All analysis where done on chromosomal replicons leaving out plasmids (specifically: pBP-B171 from BIM B-171, pSHB9 from SH-B9, pPDSLzg-1 from PDSLzg-1 and pONU554 from ONU 554). All calculations were done using stand-alone software, given that the strain MS32 had not been released by the time of the analysis, and therefore it was not possible to upload the data to web-based services. The subsequent sections describe the approach behind every tool and the insight gained from such methods. Detailed commands, arguments and further specifications of the used software is in Appendix B

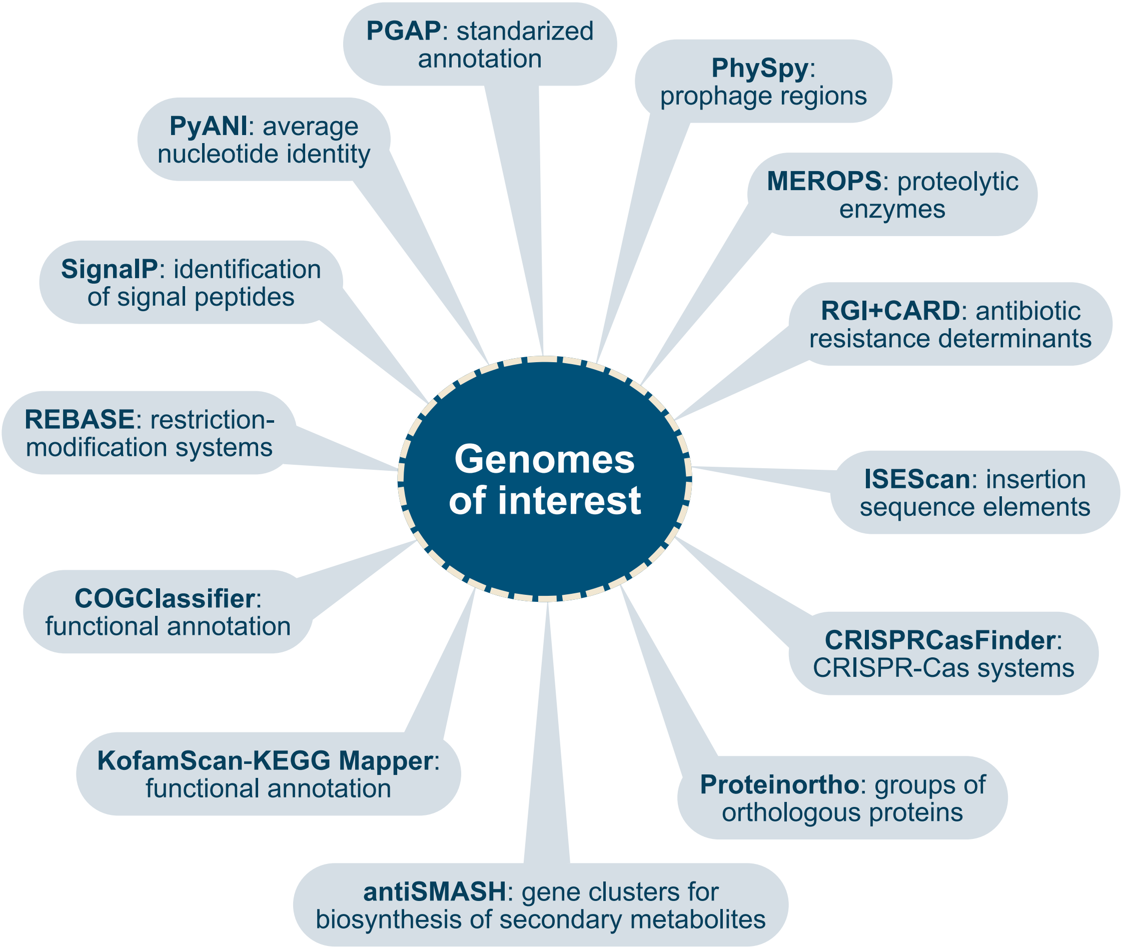

Genomes of interest
PGAP: standarized annotation
PhySpy: prophage regions
PyANI: average nucleotide identity
MEROPS: proteolytic enzymes
SignalP: identification of signal peptides
RGI+CARD: antibiotic resistance determinants
REBASE: restriction-modification systems
ISEScan: insertion sequence elements
COGClassifier: functional annotation
CRISPRCasFinder: CRISPR-Cas systems
KofamScan-KEGG Mapper: functional annotation
Proteinortho: groups of orthologous proteins
antiSMASH: gene clusters for biosynthesis of secondary metabolites

FIGURE 2.1: Graphical summary of the tools used for comparative analysis of *Bacillus* genomes.

2.2.1 Genome Annotation

Standardized annotation across samples is required before any comparative genomics experiment. Annotation is a link from sequence to biology, allowing a researcher to gain understanding of complex biological systems. As software approaches and reference databases are under constant development, predicted genomic features and their corresponding annotations might differ between genomes analyzed with different tools, and even between versions of such tools. Therefore, the genomes of this study (Table 2.1) were downloaded from the NCBI and re-annotated.

> "The value of a genome is only as good as its annotation" [304]

TABLE 2.1: *Bacillus* genomes used for comparative genome analysis. NCBI identifiers for BioSample, BioProject and Assembly are presented. Identifier NCTC10337 corresponds to DSM27.

Species	Strain	BioSample	BioProject	Assembly
B. pumilus	MS32	SAMN26309570	RJNA811128	-
B. pumilus	NCTC10337	SAMEA4076707	PRJEB6403	GCA_900186955.1
B. pumilus	145	SAMN06706381	PRJNA377620	GCA_003431975.1
B. pumilus	SH-B11	SAMN03372367	PRJNA276290	GCA_001578165.1
B. pumilus	UAMX	SAMN15498547	PRJNA645214	GCA_013423765.1
B. pumilus	BIM B-171	SAMN14228080	PRJNA770178	GCA_020535425.1
B. pumilus	SH-B9	SAMN03372270	PRJNA276289	GCA_001578205.1
B. pumilus	MTCC B6033	SAMN02677288	PRJNA239250	GCA_000590455.1
B. pumilus	150a	SAMN06706382	PRJNA377620	GCA_003571425.1
B. pumilus	TUAT1	SAMD00032095	PRJDB4002	GCA_001548215.1
B. pumilus	PDSLzg-1	SAMN05504558	PRJNA335919	GCA_001704975.1
B. pumilus	AR03	SAMN22186268	PRJNA769965	GCA_020520205.1
B. pumilus	ONU 554	SAMN15902829	PRJNA659273	GCA_014489355.1
B. pumilus	ZB201701	SAMN09215342	PRJNA471729	GCA_004006455.1
B. pumilus	SAFR-032	SAMN00253833	PRJNA20391	GCA_000017885.4
B. pumilus	EB130	SAMN20719155	PRJNA753978	GCA_019710455.1
B. subtilis	168	SAMEA3138188	PRJNA76	GCA_000009045.1
B.licheniformis	DSM13	SAMN02603292	PRJNA224116	GCF_000008425.1

Given the results of the experiments aimed to generate a germination deficient *B. pumilus* DSM27 mutant, the close relative (and already germination negative) *B. pumilus* MS32 was selected for the fermentation experiments. Since MS32 is a novel isolate, its genome was sequenced, characterized and compared with that of other *B. pumilus* strains and with *B. subtilis* 168 and *B. licheniformis* DSM13. (Chapter 6)

Prokaryotic Genome Annotation Pipeline (PGAP)

PGAP (version 2021-07-01.build5508) [312], was used for re-annotation of bacterial genomes of this study. Developed by the National Center for Biotechnology Information, U.S (NCBI), this tool allows the prediction of functional genome units such as protein-coding genes, RNAs, control regions, mobile elements etc., the approach behind PGAP consist in integrating predictions from both *ab initio* and homology-based methods (graphic pipeline summary available at `https://www.ncbi.nlm.nih.gov/genome/annotation_prok/process/` [312]. An advantage of PGAP is that annotated genomes are directly compatible with the GenBank submission system and its distribution as a container.

PyANI

PyANI is a Python software package to calculate whole genome ANI values and creates corresponding graphical outputs for easy visualization. [261]. A relevant advantage is the possibility of distribution of tasks in a multicore system, which reduces computation times in comparisons with multiple genomes. PyANI (version 0.2.1) was used on all complete genome sequences of *B. pumilus* strains.

Proteinortho

The protein sequences encoded in each *Bacillus* genome were passed to Proteinortho (version 6.0.31) [188], a software tool implementing a BLAST (Basic Alignment Search Tool) based approach to identify sets of orthologous sequences. Proteinortho is a established orthology detection tool, which performs as good as other tools such as OrthoMCL and Multi-Paranoid [239], with the advantage of being optimized for extensive datasets, efficient, and less computationally demanding by distribution of calculations over multiple processing cores [239, 188]. By this analysis it is possible to identify which proteins are unique to each strain and what are the common "core" functions encoded within *Bacillus* genomes.

antiSMASH

The first version was released in 2011, and the sixth in 2021, now describing 71 types of BGCs [30], the tool offers a low false positive rate together with a fast and comprehensive report for known BGCs [221]. Nevertheless there are some limitations, antiSMASH stands as high-confidence/low-novelty kind of approach, meaning that a potential BGCs is identified based on matching signatures from known clusters and therefore, novel or unknown types of gene clusters might be missed [221]. Analysis with antiSMASH can aid in the characterization of the secondary metabolite potential of a strain of interest.

RGI

Complementary to the CARD database, the RGI (Resistance Gene Identifier) tool [6] integrates the information at CARD to predict and annotate the resistance determinants within a genome (or metagenomic) dataset [6]. RGI has different modes for detection, "perfect" indicates exact matches within the CARD database, while "strict" algorithm allows some variation within established cut-off values, and finally, the "loose" mode identifies matches outside the cut-off scores in order to identify potential and emerging antimicrobial resistance genes. Identification of such determinants is of relevance for bacteria in industrial applications, specially in food and feed additives areas, since transmission and spread of antibiotic resistance genes is to be monitored and prevented to ensure safety of use of any strain of interest [303].

PhiSpy

This is a bioinformatic tool that achieves identification based on seven distinct characteristics of prophages, namely: protein length and similarity to known phage proteins, presence of unique phage words, transcription strand directionality, AT and GC skew values, and phage insertion sites [4]. PhiSpy (version 4.2.19) was used to scan *Bacillus* genomes and identify potential phage regions based on those characteristics. PhiSpy was also set to include a search step against the VOGdb profile hmm database (release vog210, `http://vogdb.org/`), which gathers information from all viral genomes at NCBI Refseq into curated orthologous groups.

ISEScan

ISEScan is a software pipeline that offers highly sensible and automated annotation of full-length IS [351]. One advantage of ISEScan is the identification of novel IS elements, different from the known ones in the current databases. This is achieved by a combination of strategies, first, it detects novel remote homology by implementing 621 profile hidden Markov models, therefore, it does not depend on similarity searches against the known genes in public genomic databases, and secondly it identifies inverted repeats (IR) sequences flanking the IS element by directly analyzing the input genome instead of detection based on similarity to known IR [351].

CRISPRCasFinder

This program offers a combination and upgrade of two tools, CRISPRFinder and CasFinder. It depicts enhanced performance in identification of CRISPR arrays and Cas proteins, it also offers a typing module enabling sorting according to the latest classification scheme [64].

REBASE

First launched in 1998, this database has grown into a comprehensive and curated compendium of information including: restriction enzymes, their associated methylases, commercial availability, sequence data, crystal structures, cleavage sites, recognition sequences, genome data, isochizomers and methylation sensitivity [275, 274]. To identify RM systems in the *Bacillus* genomes, a BLASTp [9, 44] search was conducted against the REBASE database (downloaded 14.02.2022).

MEROPS

The website (`https://www.ebi.ac.uk/merops/`) and accompanying database were first release in 1996 offering a classification of proteolytic enzymes into clans and families. Today, it constitutes a rich resource around peptidases and the proteins that inhibit them. It allows classification in a multilevel hierarchy from sequence to protein species, subfamily, family and clan based on structure [267, 266]. The MEROPS database (merops_scan.lib version 12.1) was used as subject in a BLAST [44, 9] search for peptidases on *Bacillus* genomes.

SignalP

With the latest version 6.0 released in 2021, SignalP implements a machine learning model to identify all five known types of signal peptides and their cleavage sites. The protein language models are sensitive enough so no additional information regarding source organism has to be provided in order to obtain accurate predictions, which is relevant for analysis of metagenomic datasets. The recent release showed improved detection performance, specially in two underrepresented types (Sec/SPIII and Tat/SPII) and improved prediction of cleavage sites [314]. SignalP is regarded as one of the most popular and user friendly programs for signal peptide prediction [37]. The five types of signal peptides identified by SignaP are:

1. Sec/SPI "standard" signal for Sec translocon and cleaved by signal peptidase (SP) I.
2. Sec/SPII signal peptide for lipoproteins trasnported by Sec and cleaved by SP II.
3. Tat/SPI signal peptide for the Tat translocon and cleaved by SPI.
4. Tat/SPII lipoprotein signal peptides exported by Tat and cleaved by SPII.
5. Sec/SPIII pilin and pilin-like signal peptides transported by the Sec pathway and cleaved by SPIII.

KofamScan and KEGG Mapper

KofamScan [13] allows to assign K numbers to protein sequences, the resulting identifiers can be loaded into KEGG Mapper (`https://www.genome.jp/kegg/mapper/reconstruct.html`) to reconstruct pathways and connect the information with other resources at KEGG. Previous tools such as BlastKOALA, GhostKOALA [164] and KAAS [230] rely on pairwise sequence comparison approaches (BLAST and GHOSTX) against KEGG to generate K number annotations. A major advantage of KofamScan is the search against profile hidden Markov models representing KO families, which is more computationally efficient. Each model is accompanied by an adaptive score threshold, the models and their scores constitute the KOfam database [13]. KofamScan (version 1.3.0 with the Kofam database as of 30.01.2022) was used to generate reliable assignments of K numbers to the proteins in the genomes of interest, the resulting annotations were visualized with KEGG Mapper.

InterProScan

This tool offers protein function classification at genome-scale by integrating information from several source databases [364, 158]. The latest release entails a modular Java-based architecture and is designed to benefit from computational cluster systems for massive scale and parallelization of computationally intensive analysis on large datasets. Output formats include TSV, XML, GFF3 and HTML files [158]. InterproScan is a robust tool employed in genome sequencing projects and by the UniProt Knowledgebase (UniProtKB [59]), here it was used to further characterize particular proteins from the *Bacillus* of interest.

COGclassifier

COGclassifier is an easy to use command line tool that provides straight forward COG functional classification of proteins of interest and creates publication ready visualizations of the results [295]. COGclassifier (version 1.2.0) was used for the analysis of *Bacillus* genomes.

2.3 Small-scale fermentations

The next step in this comparative study of *B. pumilus* against other species of *Bacillus* used as industrial enzyme producers was to select fermentation conditions that support similar growth rates for them so a comparison is possible. Variations of a rich media with and without potato dextrose broth and other additives were tested, yeast extract concentrations were also modified. Additionally, SMM media during preculture stage was evaluated. The conditions best suited for the fermentation experiments with the targeted *Bacillus* are described below. The objective of the fermentation experiments was to generate samples for transcriptomic analysis of these *Bacillus* under conditions closer to industrial productive processes. The collected samples were processed with an optimized RNA isolation protocol presented in Chapter 6.

B. pumilus MS32, *B. licheniformis* MW3 Δ *yqfD* and, *B. subtilis* 168 Δ *yqfD* were re-activated from lyophilized cultures and precultured twice during 16 and 6 hours, respectively, before inoculation of 0.5 L bioreactors. For each *Bacillus* species, fermentation runs were done by triplicate. The fermentations were carried out at the Research and Development facilities of the industrial partner AB Enzymes GmbH, Darmstadt, Germany.

Super rich fermentation media consisted of: 2% Yeast Extract, 2.5% Tryptone, 1% NaH_2PO_4 x2 H_2O, 1% Na_2HPO_4 x2 H_2O, 1% Saccharose and 0.5% Potato Extract Glucose Broth. The first preculture was done on super rich fermentation media, and the second preculture was supplemented with 1% of saccharose, both precultures had a volume of 150 ml and were incubated in 1000 ml baffled shake flasks (37 °C 180 rpm). These precultures allow the cells to adapt to the fermentation media and to start the main culture with a synchronized population of actively growing bacteria. The main cultivation on the 0,5 L bioreactor was done on the same media as the second preculture and supplemented with an antifoaming agent. The pH within the fermenter started at 6,9 and afterwards controlled to 7.10 +/-0.2 with 12.5% NH3 or 12.5% H2SO4. Aeration was set to 0.5 vvm, the stirred speed was 1200 rpm, and the temperature 37°C . Samples were taken at 2.5, 4, 7 and 19 hours after inoculation of the main culture and processed according to the method described in Chapter 6.

2.4 RNA-seq Analysis

2.4.1 RNA isolation and library preparation

The sampling protocol, optimized RNA isolation method and cDNA library preparation procedures were done for 12 *B. pumilus* MS32 samples and for 12 *B. licheniformis* MW3 Δ *yqfD* samples as described in in Chapter 6. Samples from *B. subtilis* 168 fermentations were also collected and total RNA was isolated so they can be included in future analysis.

2.4.2 Bioinformatic Analysis

Quality assessment

The quality of the RNA-seq libraries was determined with FastP [52] (version 0.20.1), with parameters for adapter detection, overrepresentation analysis and base-correction enabled, while length filtering was disabled. SortmeRNA [181] (version 4.3.3) was used to detect reads derived from rRNA that remained after the rRNA depletion step. The tin.py function of RSEQC package [335] (version 4.0.0) was used to establish the TIN (transcript integrity number) of the libraries.

Data processing pipelines

FastQ files containing reads that passed the FastP quality filters were used as input for READemption ([104], version 1.0.10) a pipeline for computational evaluation of RNA-seq data. Within READemption [104] alignment to the corresponding reference genomes, strand-specific coverage calculation, and quantification based on reference annotation were performed. Moreover, differential gene expression analysis was done with DESeq2 as implemented within the package. After the initial analysis by READemption [104], the package Enhanced Volcano Plots [29] was used to visualize the output of the differential expression analysis.

Coverage normalized by the total number of aligned reads multiplied by the lowest number of aligned reads within the considered sample set was generated by READemption [104] as wig (Wiggle) files. These files were used as input for Annogesic [363] (version 1.0.22). Annogesic is a modular command-line tool integrating different analysis for RNA-seq data, such as those for detection of: genes, CDSs, tRNAs, rRNAs, transcripts, terminators, small open reading frames and small RNAs. Additionally, Annogesic implements RNAup, RNAplex and IntaRNA for sRNA target prediction.

For sRNA identification, candidate sequences were filtered according to the energy change of the predicted secondary structure (normalized by sequence length), with the default cutoff value of -0.05. Additionally, a database of known sRNAs was provided for homology detection via Blast+ search (cutoffs for e-value= 0.0001 and score=40). The database contained sRNA matches to the RFAM [121] database and sRNA sequences reported previously in literature. The database was collected by Anton Farr during his master degree studies, and is part of a Nextflow pipeline for RNA-seq analysis focused on multi-species comparisons [88, 87]. For a sRNA candidate to be reported it had to be found in all three replicates of each data set. The gff file of sRNAs detected by Annogesic [363] was passed to READemption [104] for its quantification in order to create the raw input for clustering the transcriptional profiles. Identification of additional non-coding RNAs was done by scanning the genomes using the covariance models of the RFAM database [112, 159, 121] (version 14.7) together with the cmscan command (version 1.1.4) from the Infernal package [238] (version 1.1). Lower-scoring overlapping matches were removed, keeping only the best matching one for a given sequence region, as recommended in the RFAM documentation.

Clustering of transcriptional profiles

DP_GP_cluster [220] (version v.0.1) was used to cluster features with similar transcriptional trajectories during the fermentation. The number of clusters is determined by a Dirichlet process (DP), while gene trajectory and time dependency is calculated by a Gaussian process (GP), both in a nonparametric way [220]. To create the input for DP_GP_cluster, the quantification counts as TPM from READemption were used. First the mean transcriptional activity at each time point was determined. The values were then transformed with the hyperbolic arcsine (Asinh) function, this has the advantage, unlike log transformations, to accept zero counts. The transformed values were Z-scaled and used as input for DP_GP_cluster.

Dedicated in-house Python scripts were used to merge, condense and summarize the data generated with Annogesic, READemption and DP_GP_cluster. The scipy [331] package was used to calculate the correlation between the transcriptional activities across the fermentation time points of a predicted sRNA and its potential target.

Chapter 3

Results

3.1 A collection of germination deficient *Bacillus* strains

Transformation of *B. subtilis* 168 to generate a germination deficient mutant was achieved. A total number of colonies of 10, 21, and 12 were obtained for the reactions with 300, 500 and 750 ng of DNA, respectively, showing nicely how the DNA concentration affects the amount of mutants generated. Further PCR and sequencing confirmation experiments showed that the amplified fragments corresponded to the expected band sizes corroborated the presence of *ermD* and its correct integration within the *B. subtilis* 168 genome.

Regarding the *B. pumilus* DSM27 germination deficient mutant, none of the tested methods produced the mutant. Despite multiple attempts to optimize the protoplast transformation protocol, experiments were unsuccessful. The cells recovered their bacilliar form again, but screenings failed to identify the desired mutant. It might be the case that random mutations allowed spontaneous resistance or that the *ermD* was (even partially) incorporated elsewhere in the genome. The presence of *ermD* was not confirmed by the PCR reactions, neither the control for integration produced the desired signal, while the *yqfD* was still present in the tested clones.

Transconjugation of *B. pumilus* DSM27 was unsuccessful, as only empty plates or plates with cellular debris were obtained. This indicates that the method with the tested conditions is suitable for other bacilli such as *B. subtilis* and *B. licheniformis* but not for this evidently recalcitrant strain. This was also the case for the Tribos transformation, were the control plates without antibiotics at least confirmed that the cells survived the treatment.

Therefore, *B. pumilus* DSM27 was substituted by the strain MS32 in order to proceed with the next phase of experiments. *B. pumilus* MS32 is already genetically accessible and was kindly provided by the industrial partner AB Enzymes GmbH. Following sections will present the results of comparative genomic analysis used to characterize MS32 and sustain the selection of this strain as an adequate alternative to DSM27.

Even with a more domesticated *B. pumilus* MS32, efficient transformation was challenging. Electroporation experiments to introduce a plasmid into this strain were only successful with plasmids isolated from specific strains presumably sharing a compatible methylation profile, pointing to RM systems as a barrier for transformation.

By having a collection of *B. subtilis*, *B. licheniformis* and *B. pumilus* strains unable to produce viable spores, experiments at industrial facilities using small-scale fermentations were possible. Such experiments provided significant insights into the biology of these bacteria when exposed to industrial scenarios and how their different adaptations relate to productivity.

3.2 Comparative genomics of *Bacillus pumilus*

By combining the predictions made by each of the bioinformatic tools described, a nice overview of the main features of interest was obtained. The main results of this section of the analysis were summarized and prepared for a genome announcement publication (Chapter 6).

3.2.1 MS32 belongs to the *B. pumilus* species and is closely related to DMS27

In order to characterize the novel *B. pumilus* MS32 in regards to other *B. pumilus* strains, the bacterial genomes were analyzed with the bioinformatic tools previously described. Integrating the output generated by the diverse software tools allowed a characterization of the genomic features encoded by *B. pumilus*, as well as potential optimization points of interest for industrial applications.

The genome of *B. pumilus* MS32 consists of single circular chromosome of 3,824,664 base-pairs (bp), and presents a G+C content of 41.6% with similar features to the other *B. pumilus* strains (Table 3.1). No plasmids were identified during the genome analysis.

TABLE 3.1: Comparison of genomic features of *B. pumilus* MS32 with other *B. pumilus* strains and the closely related *B. subtilis* and *B. licheniformis*

Species	Strain	Chromosome size (bp)	G+C content (%)	Genes	CDSs	rRNAs (5S, 16S, 23S)	tRNAs	Pseudogenes
B. pumilus	MS32	3824664	41.60	3,880	3,770	8, 8, 8	81	58
B. pumilus	NCTC10337	3855667	41.71	3,969	3,859	8, 8, 8	81	61
B. pumilus	145	3937399	41.16	4,054	3,944	8, 8, 8	81	49
B. pumilus	SH-B11	3860091	41.32	3,936	3,826	8, 8, 8	81	24
B. pumilus	UAMX	3854893	41.71	3,961	3,851	8, 8, 8	81	72
B. pumilus	BIM B-171	3814325	41.67	3,945	3,835	8, 8, 8	81	37
B. pumilus	SH-B9	3787586	41.57	3,876	3,766	8, 8, 8	81	36
B. pumilus	MTCC B6033	3763493	41.37	3,859	3,749	8, 8, 8	81	37
B. pumilus	150a	3747740	41.35	3,789	3,678	8, 8, 8	82	54
B. pumilus	TUAT1	3723433	41.42	3,838	3,728	8, 8, 8	81	19
B. pumilus	PDSLzg-1	3698973	41.96	3,778	3,668	8, 8, 8	81	43
B. pumilus	AR03	3655835	41.83	3,718	3,608	8, 8, 8	81	54
B. pumilus	ONU 554	3642544	41.93	3,708	3,598	8, 8, 8	81	54
B. pumilus	ZB201701	3640542	41.86	3,701	3,591	8, 8, 8	81	43
B. pumilus	SAFR-032	3704641	41.29	3,741	3,643	7, 7, 7	72	61
B. pumilus	EB130	3614840	41.85	3,635	3,526	8, 8, 8	80	277
B. subtilis	168	4215606	43.51	4,407	4,286	10, 10, 10	86	42
B.licheniformis	DMS13	4222645	46.19	4,329	4,231	7, 7, 7	72	67

The average nucleotide identity (ANI) values calculated with PyANI [261] confidently placed the strain MS32 within the *B. pumilus* species (Figure 3.1). As shown in the heatmap, the strain BIM B-171 is the closest to the type strain DSM27 (NCTC10337), with an ANI of 98.69%, however it was made public at NCBI only until late 2021, and therefore was not available by the time of the fermentation experiments (the first round of analysis was done with *B. pumilus* strains available at NCBI by 2019). Nevertheless, the ANI between MS32 and DSM27 is 97.66%, being the second best candidate to substitute DSM27 for the fermentation experiments.

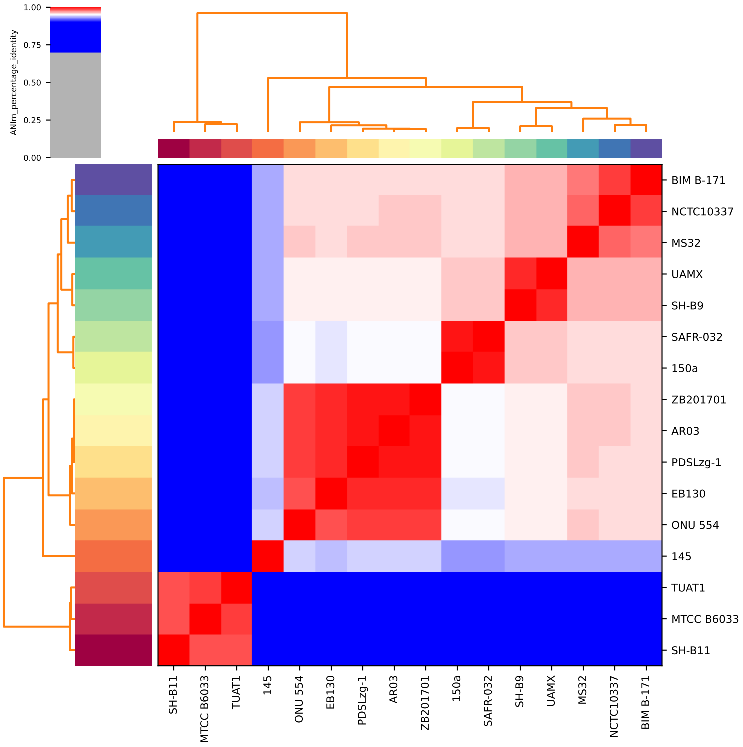


FIGURE 3.1: Average Nucleotide Identity (ANI) analysis by PyANI [261] performed for *B. pumilus* strains. NCTC10337 corresponds to the type strain DSM27.

Figure 3.2 presents the whole genome alignment of *B. pumilus* MS32 with the type strain DSM27 and the close relative BIM B-171. This kind of visual inspection of genomic data allows to identify genomic rearrangements associated to horizontal gene transfer, duplication, recombination, translocation and deletion events, which are relevant to compare genomes of interest and unveil its evolutionary story [71].

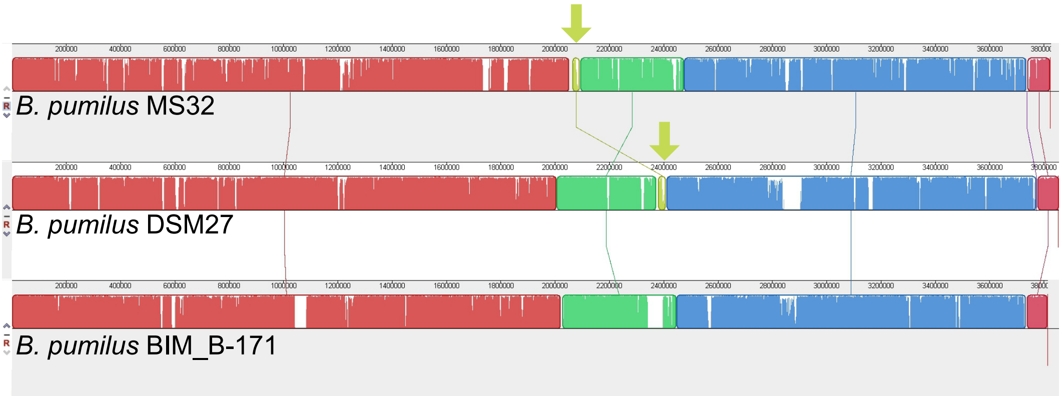


FIGURE 3.2: Genomic alignment of *B. pumilus* MS32 with the type strain DSM27 and the close relative BIM B-171, the comparison was done with progressive Mauve [71]. Putative prophage region identified with PhiSpy [4] is indicated by an arrow.

Whole genome alignment of *B. pumilus* MS32 with its closest relatives evidences a widespread co-linearity between strains (Figure 3.2). The exception being the locally collinear block from 2063898-2093164 nt in MS32, which has a different position in the type strain DSM27 and is absent in BIM B-171. According to the PhiSpy [4] analysis, this region is within a predicted prophage which ranges from 2053152nt to 2095050nt in MS32 while in DSM27 the prophage location is 2374962-2401515nt.

3.2.2 *B. pumilus* MS32 has a highly dynamic genome

The analysis done with ISEScan [351] identified 37 IS encoded by MS32, which represents a 1.32% of the genome. Sequences of the IS1182 and IS3 families were identified, with 11 and 26 members, respectively (Table 3.2). By this, MS32 is the strain with more IS elements of the analyzed bacteria, followed by the strain UAMX (24 IS). On the other hand, no IS were found for the BIM B-171, SH-B9, TUAT1, PDSLzg-1 and ARO3 strains, and one novel IS was predicted for the strain SAFR-032. However, this predictions must be taken cautiously, given the multi-copy and repetitive nature of IS elements, these regions might not be solved properly in a genome assembly that relies only in short read data, therefore the *B. pumilus* strains might carry a different number of IS depending on the sequencing technology approach implemented. As the MS32 genome was produced with an hybrid assembly combining short and long read data, the IS elements and their copies could be accurately assigned within the genome.

TABLE 3.2: Summary of predicted features within *B. pumilus* genomes. Analysis of prophage content, Insertion Sequence elements, CRISPR-Cas systems, and resistance determinants. CAT stands for Type A chloramphenicol O-acetyltransferase and Beta-lactamase refers to BPU-1 family class D beta-lactamase.

		Insertion Sequence Elements				**CRISPR/Cas sytems**		**Resistance Genes**	
***B. pumilus* strain**	**No. Prophages**	**IS1182**	**IS3**	**New IS**	**Total IS**	**CRISPR**	**Cas**	**CAT**	**Beta-lactamase**
MS32	2	11	26	0	37	1	0	+	-
NCTC10337	2	2	1	0	3	0	0	+	-
145	4	7	6	0	13	0	0	+	-
SH-B11	3	14	4	0	18	1	0	+	-
UAMX	4	18	6	0	24	0	0	+	+
BIM B-171	3	0	0	0	0	0	0	+	+
SH-B9	2	0	0	0	0	0	0	+	+
MTCC B6033	2	4	6	0	10	0	0	+	-
150a	2	5	9	0	14	0	0	+	+
TUAT1	2	0	0	0	0	0	0	+	-
PDSLzg-1	3	0	0	0	0	5	1	+	+
AR03	1	0	0	0	0	8	1	+	+
ONU 554	1	4	0	0	4	1	0	+	-
ZB201701	1	7	0	0	7	0	0	+	+
SAFR-032	2	13	2	1	16	0	0	+	+
EB130	1	3	0	0	3	0	0	-	+

The table 3.2 also presents the prophage predictions done with PhiSpy [4] on the *B. pumilus* genomes. One to four prophages were identified in the strains. MS32 carries two prophages, positioned from 2053152 to 2095050 and from 2839776 to 2885424, respectively.

Regarding CRISPR-Cas systems, there is one CRISPR array predicted for MS32 with no associated *Cas* gene identified. CRISPRCasFinder [64] assigns an evidence level of 1 to this prediction, the lowest in their scoring system, which is usually associated to short candidates that likely do not correspond to CRISPRs. Based on the analysis, CRISPR-Cas systems are scarce within the analyzed *B. pumilus* genomes (Table 3.2) as they are absent in 11 of the strains or identified with a low confidence level. There are two exceptions, the strains AR03 and PDSLzg-1, with some predictions ranking as level 3 and 4, according to the CRISPRCasFinder documentation, this rankings can be considered as highly likely candidates.

Concerning Restriction-Modification systems, a BLAST search against the REBASE [274] database returned abundant matches across the *B. pumilus* genomes, particularly from the Type II category (Table 3.3).

TABLE 3.3: Summary of matches from *B. pumilus* genomes against the REBASE database

REBASE match	MS32	NCTC 10337	145	SH-B11	UAMX	BIM B-171	SH-B9	MTCC B6033	150a	TUAT1	PDSLzg-1	AR03	ONU 554	ZB201701	SAFR -032	EB130
Type I methyltransferase	9	2	2	5	0	1	2	8	0	1	1	1	2	2	3	2
Type I specificity subunit	2	1	2	2	1	1	3	2	1	3	1	1	3	2	1	4
Type II methyltransferase	12	14	15	12	11	13	13	10	11	10	14	11	11	10	10	10
Type II nicking endonuclease	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
homing endonuclease	3	5	4	6	4	5	4	3	4	3	3	3	4	3	3	2
orphan methyltransferase	2	2	2	2	3	2	2	2	2	2	3	2	2	2	2	2
P. Type I methyltransferase	9	2	2	5	0	1	2	7	0	1	1	1	2	2	3	2
P. Type I restriction enzyme	1	0	1	1	0	0	1	1	0	1	0	0	1	1	0	1
P. Type I specificity subunit	2	1	2	2	1	1	3	1	1	3	1	1	3	2	1	4
P. Type II methyltransferase	11	13	14	11	11	12	12	9	10	9	13	10	10	9	9	9
P. Type II nicking endonuclease	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
P. Type II restriction enzyme	7	4	5	4	6	4	5	5	4	4	5	4	5	4	4	4
P. Type IIG restriction enzyme/methyltransferase	0	1	1	0	1	0	0	0	2	0	0	0	0	0	0	0
P. Type III methyltransferase	0	2	2	0	0	0	0	0	0	0	0	0	0	0	0	0
P. Type III restriction enzyme	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0
P. Type IV methyl-directed restriction enzyme	1	0	1	0	0	1	1	0	1	1	1	1	0	2	1	2
P. homing endonuclease	3	5	4	5	4	5	4	3	4	3	3	3	4	3	3	2
P. orphan methyltransferase	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0

3.2.3 Specialized metabolites and antibiotic resistance in MS32

The analysis done with RGI [6] against the CARD database found a Type A chloramphenicol O-acetyltransferase (CAT) widely spread within the *B. pumilus* strains, missing only in the EB130 genome. On the other hand, a beta-lactamase of the BPU-1 family was absent in MS32, but found in nine of the other strains, the CARD database [6] accession for this gene family is ARO:3004759 and it appears restricted to *B. pumilus* strains (Table 3.2).

Next to antibiotic resistance determinants, the faculty to produce secondary metabolites should also be assessed in strains of industrial interest. The tool antiSMASH [30] predicted 13 putative biosynthetic gene clusters (BGC) for secondary metabolites encoded by the MS32 genome (Table 3.4). Seven of those BGCs show some similarity degree to previously known clusters, while six appear to be novel.

TABLE 3.4: Biosynthetic gene clusters found in *B. pumilus* MS32

Region	Type	From	To	Most similar known cluster	Similarity
1	NRPS	348933	431113	lichenysin (NRP)	85%
2	lanthipeptide-class-iii	544454	567141		
3	NRPS,T1PKS	666747	746768	zwittermicin A (NRP + Polyketide)	18%
4	RRE-containing	930317	949691		
5	terpene,siderophore	1099536	1128153	carotenoid (Terpene)	50%
6	RRE-containing,LAP	1638436	1661599	plantazolicin (RiPP:LAP)	100%
7	betalactone	1869668	1896496	fengycin (NRP)	53%
8	terpene	1959128	1981002		
9	T3PKS	2019225	2060325		
10	betalactone	2567250	2599699		
11	RiPP-like	3424811	3435101		
12	other	3467958	3509379	bacilysin (Other)	85%
13	NRPS	3760875	3808023	bacillibactin (NRP)	53%

3.2.4 Unique proteins of *B. pumilus* MS32

The genome of *B. pumilus* MS32 was predicted to encode 3712 proteins, 3382 with annotation and 330 designated as "hypothetical protein". The analysis with Proteinortho [188] showed that MS32 shares 3489 groups of orthologous proteins with the type strain DSM27, once more confirming these two strains as very similar and MS32 as a suitable candidate to replace the type strain DSM27 in fermentation and transcriptomic experiments. Table 3.5 presents proteins unique to *B. pumilus* MS32.

TABLE 3.5: Set of proteins found in *B. pumilus* MS32 with no shared orthology to proteins encoded by other *B. pumilus* genomes according to Proteinortho [188] analysis.

LocusTag	Product
BP32_000536	class III lanthipeptide
BP32_000537	class III lanthipeptide
BP32_000538	serine protease
BP32_000539	class III lanthionine synthetase LanKC
BP32_000540	class III lanthipeptide
BP32_000541	class III lanthipeptide
BP32_000542	class III lanthipeptide

Table 3.5 continued from previous page

BP32_000543	class III lanthipeptide
BP32_000624	restriction endonuclease subunit S
BP32_000625	AAA family ATPase
BP32_000627	restriction endonuclease
BP32_001170	hypothetical protein
BP32_001171	hypothetical protein
BP32_001172	hypothetical protein
BP32_001173	hypothetical protein
BP32_001174	ETX/MTX2 family pore-forming toxin
BP32_001175	DUF3102 domain-containing protein
BP32_001179	hypothetical protein
BP32_001722	hypothetical protein
BP32_001736	collagen-like protein
BP32_001790	hypothetical protein
BP32_001791	hypothetical protein
BP32_001792	hypothetical protein
BP32_001793	hypothetical protein
BP32_001794	DGQHR domain-containing protein
BP32_001795	helix-turn-helix transcriptional regulator
BP32_001797	TniQ family protein
BP32_001798	ATP-binding protein
BP32_001799	DDE-type integrase/transposase/recombinase
BP32_001800	hypothetical protein
BP32_001801	DEAD/DEAH box helicase
BP32_001879	hypothetical protein
BP32_001880	hypothetical protein
BP32_001881	hypothetical protein
BP32_001882	hypothetical protein
BP32_001883	hypothetical protein
BP32_001884	hypothetical protein
BP32_001885	hypothetical protein
BP32_001886	hypothetical protein
BP32_002034	hypothetical protein
BP32_002035	hypothetical protein
BP32_002065	type II toxin-antitoxin system PemK/MazF family toxin
BP32_002068	XRE family transcriptional regulator
BP32_002069	hypothetical protein
BP32_002074	hypothetical protein
BP32_002084	hypothetical protein
BP32_002087	recombinase family protein
BP32_002227	hypothetical protein
BP32_002499	YjcZ family sporulation protein
BP32_002751	collagen-like protein
BP32_002763	YtxH domain-containing protein
BP32_002901	hypothetical protein
BP32_002971	hypothetical protein
BP32_002972	hypothetical protein
BP32_002973	hypothetical protein
BP32_003616	hypothetical protein
BP32_003618	hypothetical protein
BP32_003680	mannonate dehydratase
BP32_003719	MBL fold metallo-hydrolase
BP32_003748	acyltransferase family protein
BP32_003761	CDP-glycerol glycerophosphotransferase family protein

3.2.5 Comparison of *B. pumilus* MS32 with *B. subtilis* and *B. licheniformis*

General Genomic Features

From the table 3.1 some inter-species differences between *B. pumilus* MS32, *B. licheniformis* DSM13 and *B. subtilis* 168 are already evident. Starting with chromosome size and GC content, the comparison showed that genomes of *B. pumilus* are smaller and elicit lower in GC% than the *B.subtilis* and *B. licheniformis* counterparts. The GC content of *B. pumilus* (41.6%) is lower than that of *B. subtilis* 168 (43.51%) and *B. licheniformis* DSM13 (46.19%). Other differences between the three *Bacillus* of interest are the number of tRNAs and rRNAs. In these aspects, *B. pumilus* MS32 has intermediate values between *B. subtilis* 168 and *B. licheniformis* DSM13 (Table 3.1).

Protein orthology

Regarding proteins encoded by the three *Bacillus* species of interest, Proteinortho [188] was used to identify groups of orthologous proteins. Figure 3.3 presents a Venn diagram comparing groups of orthologous proteins and singletons of *B. pumilus* MS32, *B. subtilis* 168 and *B. licheniformis* DSM13.

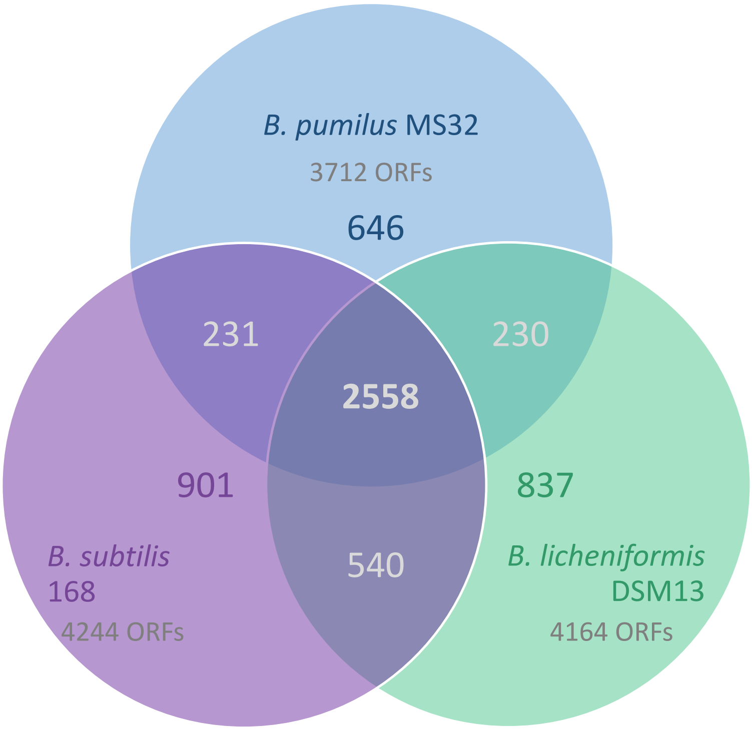


FIGURE 3.3: Venn diagram comparing groups of orthologous proteins and singletons of *B. pumilus* MS32, *B. subtilis* 168 and *B. licheniformis* DSM13. For purposes of this graph, orthologous groups are treated as entities and depicted in overlapping areas. Total ORFs are indicated in gray.

The figure 3.3 shows 2558 groups shared between *B. pumilus* MS32, *B. licheniformis* DSM13 and *B. subtilis* 168. MS32 encodes 646 predicted proteins with no identified orthologous in the other species, while similar orthologous are found in pairwise comparisons with the other two species. Out of the unique sequences of MS32, 187 are annotated as hypothetical proteins.

Antibiotic resistance

Concerning antibiotic resistance determinants, *B. subtilis* 168 is predicted to encode 11 genes related to resistance mechanisms, mostly antibiotic efflux pumps of the major facilitator superfamily (MFS) and small multidrug resistance types of transporters. Whereas no match was found for *B. licheniformis* DSM13. However, a recent study comparing more than 100 *B. licheniformis* and *B. paralicheniformis* genomes identified putative *cat* (chloramphenicol), *aph* - *aadK* (streptomycin) and *ermD* (erythromycin) resistance genes as part of an ancient resistome intrinsic to these *Bacillus* and more distantly related to other characterized instances of these genes. [3].

KEGG functional Annotation

KofamScan [13] was used to assign K numbers to the proteins encoded by *B. pumilus* MS32, in total 2565 K identifiers were assigned. The subset of locus_tags for proteins uniquely encoded by MS32 together with their K numbers were uploaded to KEGG mapper and summarized via the BRITE hierarchy (Table 3.6). The most abundant elements of the set were enzymes related to metabolism. Among those, matches for oxidoreductases, transferases, hydrolases, lyases, isomerases, ligases and translocases were identified. Within the transcription factor category, several members of families presenting a Helix-turn-helix signature were found, as well as members for the BglG family of transcriptional antiterminators. Regarding transporters, proteins for translocation of substrates such as osmoprotectants, glucose/mannose, L-cystine, iron and galactitiol were identified, including ABC transporters, Major facilitator superfamily (MFS) and Phosphotransferase (PTS) systems.

TABLE 3.6: BRITE functional hierarchy for proteins unique to *B. pumilus* MS32 based on KEGG [163] functional assignment.

KEGG identifier, module descriptions and orthologs		
ko00001	KEGG Orthology (KO)	179
Protein families: metabolism		
ko01000	Enzymes	77
ko01001	Protein kinases	3
ko01009	Protein phosphatases and associated proteins	1
ko01002	Peptidases and inhibitors	6
ko01003	Glycosyltransferases	1
ko01011	Peptidoglycan biosynthesis and degradation proteins	5
ko01004	Lipid biosynthesis proteins	1
ko01008	Polyketide biosynthesis proteins	4
ko01007	Amino acid related enzymes	1
Protein families: genetic information processing		
ko03000	Transcription factors	21
ko03021	Transcription machinery	1
ko03009	Ribosome biogenesis	1
ko03016	Transfer RNA biogenesis	2
ko03110	Chaperones and folding catalysts	1
ko03032	DNA replication proteins	2
ko03036	Chromosome and associated proteins	2
ko03400	DNA repair and recombination proteins	1
Protein families: signaling and cellular processes		
ko02000	Transporters	45
ko02044	Secretion system	1
ko02022	Two-component system	5
ko02035	Bacterial motility proteins	1
ko04147	Exosome	2
ko02048	Prokaryotic defense system	7
ko01504	Antimicrobial resistance genes	2
ko00537	Glycosylphosphatidylinositol (GPI)-anchored proteins	1

To have a general overview of common and unique pathways present in the three *Bacillus* species of interest, the K number identifiers for each genome were uploaded to KEGG mapper and a global map of metabolic pathways was generated. Figure 3.4 and 3.5 represent a comparison of pathways identified in *B. pumilus* MS32 against *B. subtilis* 168 and *B. licheniformis* DSM13, respectively. In each image, common elements are represented in blue while green corresponds to features uniquely identified in *B. pumilus* MS32. In both figures a clear overlap between the major parts of the map is evidenced.

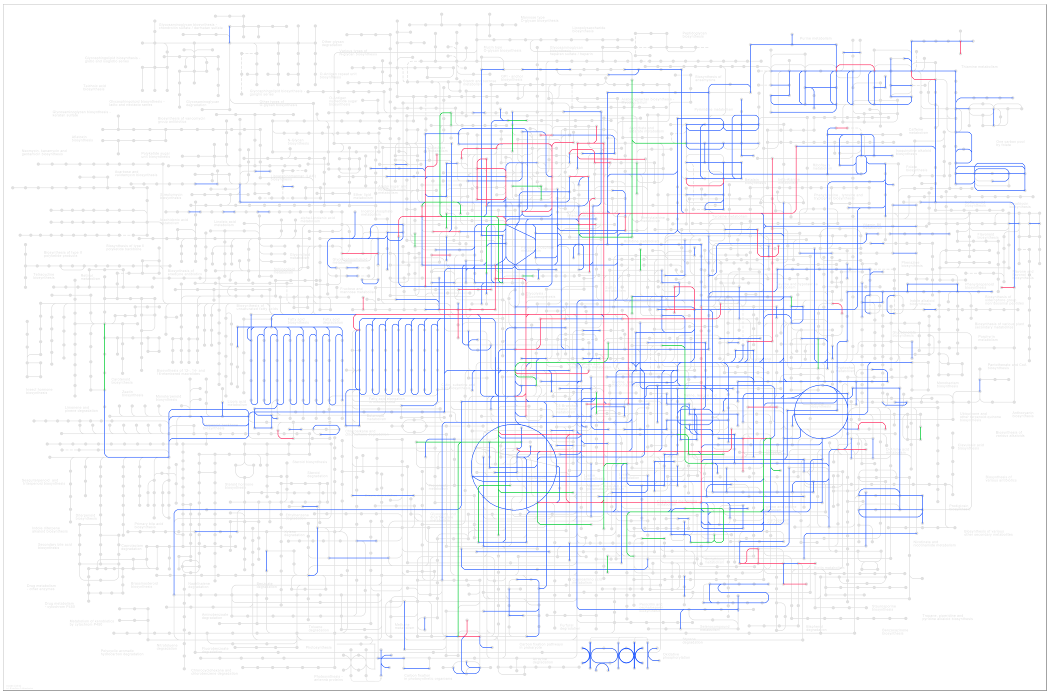


FIGURE 3.4: KEGG [163] metabolic pathway map comparison between annotated features of *B. pumilus* and *B. subtilis*. Blue=common elements, Green=unique to *B. pumilus*, Red=unique to *B. subtilis*.

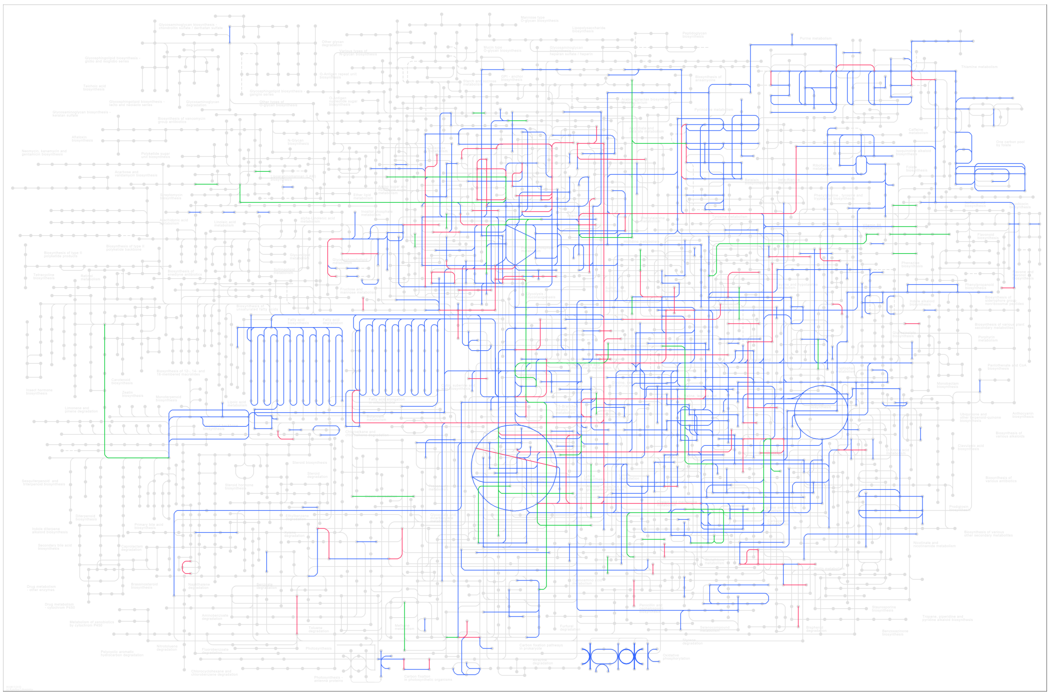


FIGURE 3.5: KEGG [163] metabolic pathway map comparison between annotated features of *B. pumilus* and *B. licheniformis*. Blue=common elements, Green=unique to *B. pumilus*, Red=unique to *B. licheniformis*.

COG functional assignment

COGClassifier [295] was used to associate a functional category to the proteins encoded by the *Bacillus* genomes according to the Cluster of Orthologous Genes (COG) database [311, 111, 310]. Table 3.7 presents the absolute counts and the corresponding percentages (relative to the total protein sequences encoded per genome) for each category. Figure 3.6 depicts a circular representation of *B. pumilus* and its genomic features together with conserved CDS in *B. licheniformis* and *B. subtilis* and colorized according to COG functional classification.

TABLE 3.7: COG functional classification for the proteins encoded by *B. pumilus*, *B. licheniformis* and *B. subtilis* genomes. Absolute counts and percentage relative to proteome size are presented.

		B. pumilus MS32		*B. licheniformis DSM13*		*B. subtilis 168*	
	COG Category	**Count**	**%**	**Count**	**%**	**Count**	**%**
J	Translation, ribosomal structure and biogenesis	248	6.68	250	6.00	247	5.82
A	RNA processing and modification	0	0.00	0	0.00	0	0.00
K	Transcription	269	7.25	288	6.92	279	6.57
L	Replication, recombination and repair	107	2.88	119	2.86	137	3.23
B	Chromatin structure and dynamics	0	0.00	0	0.00	0	0.00
D	Cell cycle control, cell division, chromosome partitioning	72	1.94	77	1.85	80	1.89
Y	Nuclear structure	0	0.00	0	0.00	0	0.00
V	Defense mechanisms	78	2.10	114	2.74	87	2.05
T	Signal transduction mechanisms	162	4.36	171	4.11	177	4.17
M	Cell wall/membrane/envelope biogenesis	173	4.66	184	4.42	211	4.97
N	Cell motility	43	1.16	42	1.01	41	0.97
Z	Cytoskeleton	0	0.00	2	0.05	2	0.05
W	Extracellular structures	0	0.00	0	0.00	0	0.00
U	Intracellular trafficking, secretion, and vesicular transport	22	0.59	24	0.58	27	0.64
O	Posttranslational modification, protein turnover, chaperones	109	2.94	124	2.98	120	2.83
X	Mobilome: prophages, transposons	58	1.56	46	1.10	29	0.68
C	Energy production and conversion	151	4.07	176	4.23	168	3.96
G	Carbohydrate transport and metabolism	235	6.33	326	7.83	295	6.95
E	Amino acid transport and metabolism	293	7.89	305	7.32	281	6.62
F	Nucleotide transport and metabolism	87	2.34	94	2.26	103	2.43
H	Coenzyme transport and metabolism	165	4.45	168	4.03	185	4.36
I	Lipid transport and metabolism	130	3.50	138	3.31	141	3.32
P	Inorganic ion transport and metabolism	138	3.72	161	3.87	167	3.93
Q	Secondary metabolites biosynthesis, transport and catabolism	55	1.48	56	1.34	70	1.65
R	General function prediction only	174	4.69	195	4.68	199	4.69
S	Function unknown	142	3.83	170	4.08	166	3.91
	Total sequences classified	2911	78.42	3230	77.57	3212	75.68

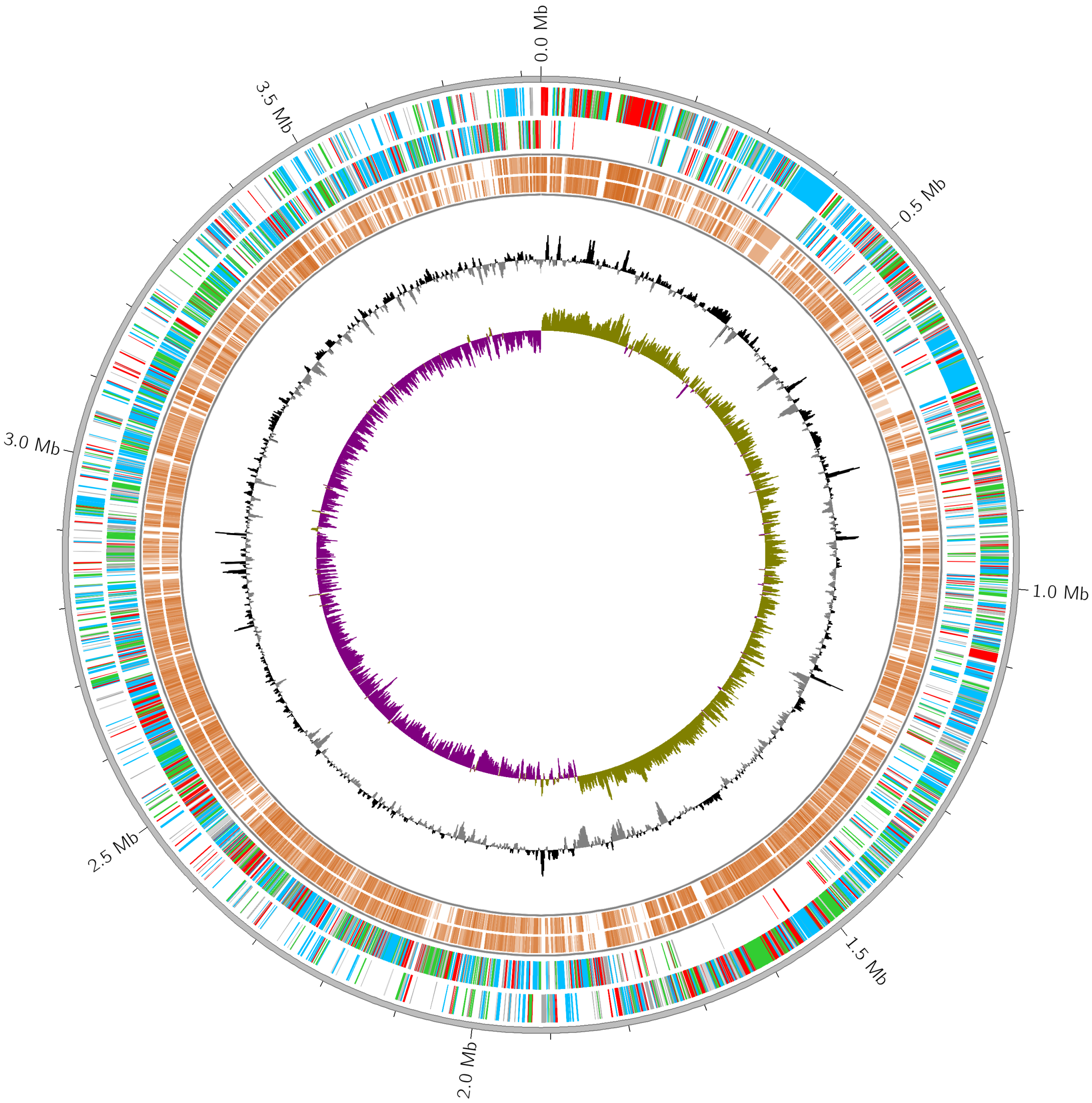


FIGURE 3.6: Circular genome plot of *B. pumilus* MS32 and conserved CDS features in *B. licheniformis* DSM13 and *B. subtilis* 168. Inner to outer tracks: 1-GC skew, 2-GC content, 3-Forward and Reverse CDS 4-Conserved CDS of *B. subtilis* 168, 5-Conserved CDS of *B. licheniformis* DSM13. COG functional categories colorized as follows: Information storage and processing (J,A,K,L,B) = red, Cellular processes and signaling (D,Y,V,T,M,N,Z,W,U,O,X) = limegreen, Metabolism (C,G,E,F,H,I,P,Q) = deepskyblue, Poorly characterized (R,S) = lightgrey, No COG classification = darkgrey. Created with MGCplotter [296], which determines conserved CDS by MMseqs2 RBH method.

For *B. pumilus* MS32 a functional category was assigned to 78.42% of the input sequences, which is comparable to the 77.57% and 75.68% of coverage obtained for *B. licheniformis* DSM13 and *B. subtilis* 168, respectively (Table 3.7). Despite the differences in genome size, the three species share fairly similar relative percentages of encoded proteins dedicated to each functional category. However, there are some differences, for example the amount of proteins associated to translation, transcription and replication functions (J, K, L categories) was higher in *B. pumilus* with 16.81% than in *B. licheniformis* and *B. subtilis* with 15.78% and 15.62%, respectively.

Additionally, "T: Signal transduction" and "X: Mobilome" categories were also slightly higher for *B. pumilus* (Table 3.7). The latter fits nicely to previous observations regarding the dynamic components found within the genome.

Notably, there is an interesting shift regarding transport and metabolism of nutrients between the species. *B. pumilus* showed a higher percentage of proteins associated to "E: Amino acid transport and metabolism" (7.89%) than *B. subtilis* and *B. licheniformis* (Table 3.7). This value is also higher than the percentage assigned to "G: Carbohydrate transport and metabolism" in *B. pumilus*. The trend is also true for the other *B. pumilus* strains, and it is inverted in *B. subtilis* and *B. licheniformis*, were category G is prevalent over E. Regarding the percentage of proteins dedicated to "C: Energy production and conversion", *B. pumilus* has intermediate values between *B. licheniformis* and *B. subtilis*.

The "E: Amino acid transport and metabolism" category, considering only proteins without homology identified by Proteinortho [188] was further analyzed. Figure 3.7 depicts a tag cloud visualization with the more prominent terms associated to the COG annotation for these proteins (generated by the online tool `https://www.freewordcloudgenerator.com/`).

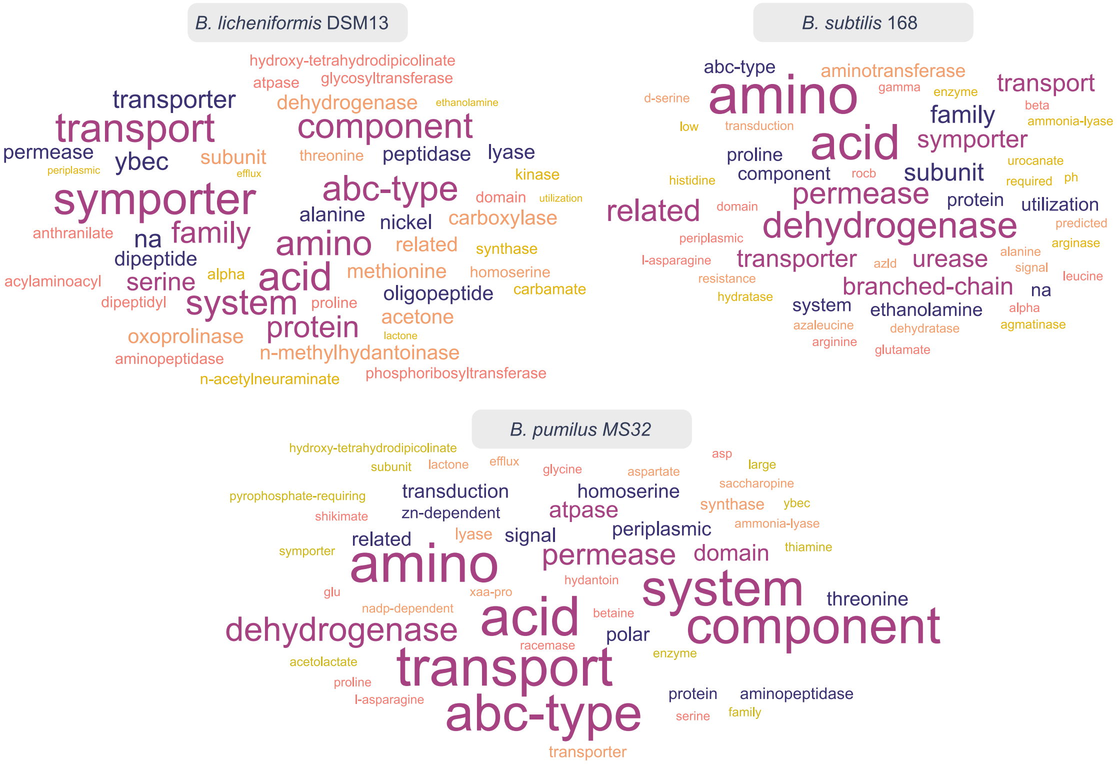


FIGURE 3.7: Tag cloud representing the most abundant terms in the COG annotation ("E: Amino acid transport and metabolism" category) for proteins without detected homology between the *Bacillus* genomes.

The Tag cloud for *B. pumilus* MS3 revealed that ABC-type transporters are enriched within this bacteria. Within the E category subset of sequences unique to each species, twelve proteins with an "ABC-type" related annotation were identified for MS32, whereas six for *B. licheniformis* and two for *B. subtilis*. At the genome scale 184, 201 and 188 proteins related to ABC systems were identified for *B. pumilus*, *B. licheniformis* and *B. subtilis*, respectively. Considering the smaller genome size of *B. pumilus*, this represents again a higher proportion of proteins dedicated to these transport systems, such abundance could facilitate more efficient nutrient uptake for *B. pumilus*.

Mobile genetic elements and CRISPRCas systems

Unlike *B. pumilus* MS32, there are no IS predicted for *B. subtilis* 168, while *B. licheniformis* DSM13 encodes 11 IS (9 from the IS3 family and 2 identified as new). Prophages are more abundant in DSM13 and 168, with 4 and 3 predictions, respectively. No CRISPR-Cas was detected in *B. licheniformis*, while three arrays were identified in *B. subtilis*, however CRISPRCasFinder assigns a low level of confidence for this prediction and no Cas protein was found.

Secretory systems

Secretion capacity is one of the most valued features of *Bacillus* strains used as industrial production platforms. Therefore, the list of genes related to this function in *B. subtilis* was retrieved from SubtiWiki [250] (Category 3.3.5: Protein Secretion, Table 3.8) and Proteinortho [188] was used to identify orthologous products encoded by the other *Bacillus* strains. The set consisted of 52 proteins, including elements from the Sec and TAT secretory pathways, flagellar export apparatus, molecular chaperons, signal peptidases and regulators. For every element in the list, the corresponding ortholog was found in *B. pumilus* MS32 and in *B. licheniformis* DSM13. The secretory machinery is a conserved feature of this bacteria and MS32 has the elements present in closely related *Bacillus* already known for their superior secretion capabilities, pointing *B. pumilus* as a good candidate for production of extracellular products.

TABLE 3.8: Genes related to the secretory machinery in *B. subtilis* and their functions. Retrieved from SubtiWiki [250]

Name	Function
comC	genetic competence
csaA	protein secretion
dnaK	protein quality control
ecsA	regulation of the secretion apparatus and of intra-membrane proteolysis
ecsB	regulation of the secretion apparatus and of intra-membrane proteolysis
ffh	presecretory protein translocation
flhA	flagellum and nanotube assembly
flhB	flagellum and nanotube assembly
fliP	flagellum and nanotube assembly
fliQ	flagellum and nanotube assembly
fliR	flagellum and nanotube assembly
fliZ	flagellum and nanotube assembly
ftsY	protein secretion
liaH	resistance against oxidative stress and cell wall antibiotics, protein secretion
lspA	protein secretion
mstX	unknown
prsA	protein folding
rasP	control of cell division, and SigV and SigW activity
rnc	processing and degradation of RNA molecules
scr	presecretory protein translocation
secA	protein secretion
secDF	protein secretion
secE	protein secretion
secG	protein secretion
secY	protein secretion
sipS	protein secretion
sipT	protein secretion
sipU	protein secretion
sipV	protein secretion
sipW	biofilm formation
spoIIIAB	activation of SigG
spoIIIAC	activation of SigG
spoIIIAD	activation of SigG
spoIIIAE	activation of SigG
spoIIIAF	activation of SigG
spoIIIAG	activation of SigG
spoIIIAH	activation of SigG, forespore encasement by the spore coat
tatAC	TAT protein secretion
tatAD	TAT protein secretion
tatAY	TAT protein secretion
tatCD	TAT protein secretion
tatCY	TAT protein secretion
yhcS	anchoring of proteins to the cell wall
yidC1	membrane insertion of proteins and protein secretion
yidC2	membrane insertion of proteins and protein secretion
ylxM	presecretory protein translocation
yrbF	protein secretion
yueB	export of YukE
yueC	export of YukE
yukB	export of YukE
yukC	export of YukE
yukD	export of YukE

Signal Peptides

Prediction of signal peptides was carried out by the updated version of SignalP software [314]. Table 3.9 presents the identified signal peptides in *B. pumilus* MS32, *B. licheniformis* DSM13 and *B. subtilis* 168.

TABLE 3.9: Signal Peptide predictions for *B. pumilus*, *B. licheniformis* and *B. subtilis*

Type of Signal Peptide	*B. pumilus* MS32	*B. licheniformis* DSM13	*B.subtilis* 168
SP	144	150	158
LIPO	98	109	112
TAT	2	3	4
PILIN	3	3	3
Total	247	265	277

The most abundant types of signal peptide for the *Bacillus* of interest were "SP" and "LIPO", which target proteins to be exported by the Sec pathway 3.9. Even though the amount of proteins predicted to carry an export signal in *B. pumilus* MS32 is lower than that of the other *Bacillus*, it represents a slightly higher percentage of the total proteins produced by the strain (6.56%) in contrast to *B. licheniformis* and *B. subtilis* with 6.36% and 6.53%, respectively.

Proteases

The MEROPS database [267, 266] was used to identify proteases encoded by the *Bacillus* species of interest. Table 3.10 depicts the corresponding results grouped by proteolytic enzyme family.

TABLE 3.10: Predicted proteases of *Bacillus* species according to the MEROPS database [267, 266].

Proteolytic Enzyme Family	*B. pumilus* MS32	*B. licheniformis* DSM13	*B. subtilis* 168
Aspartic	2	3	3
Cysteine	21	24	25
Glutamic	0	0	1
Metallo	41	48	56
Asparangine	1	1	1
Mixed	0	0	0
Serine	45	55	50
Threonine	3	4	4
Unknown	3	3	3
Total	116	138	143
Inhibitors	2	3	2

The most abundant predictions for the three species were members of the Serine protease family, with 55 proteins in *B. licheniformis* DSM13, 50 in *B. subtilis* 168, and 45 in *B. pumilus* MS32. *B. subtilis* is predicted to encode the highest amount of proteases, 143, while *B. licheniformis* and *B. pumilus* have 138 and 116, respectively.

Despite these differences, the amount of secreted peptidases is similar for the three species. Sec signal peptides were predicted for 28 proteases in *B. subtilis*, 27 in *B. licheniformis*, and 26 for *B. pumilus*.

3.3 Fermentation sampling points

A set of small-scale fermentations in rich media supporting high-cell density were implemented for *B. pumilus* MS32, *B. licheniformis* MW3 and *B. subtilis* 168 Δ *yqfD*. The aim of such experiments was to generate comparable samples from these three species in order to proceed with comparative transcriptomic analysis. The selected sampling points allowed a comparison between different growth phases in a close to production fermentation environment. Samples from *B. subtilis* were not sequenced but high-quality total RNA was prepared and could be used for additional comparative studies. Figure 3.8 presents the carbon dioxide parameters during the fermentations, this is taken to indirectly monitor and infer the culture growth status, as the turbidity of the media does not allow accurate optical density measurements.

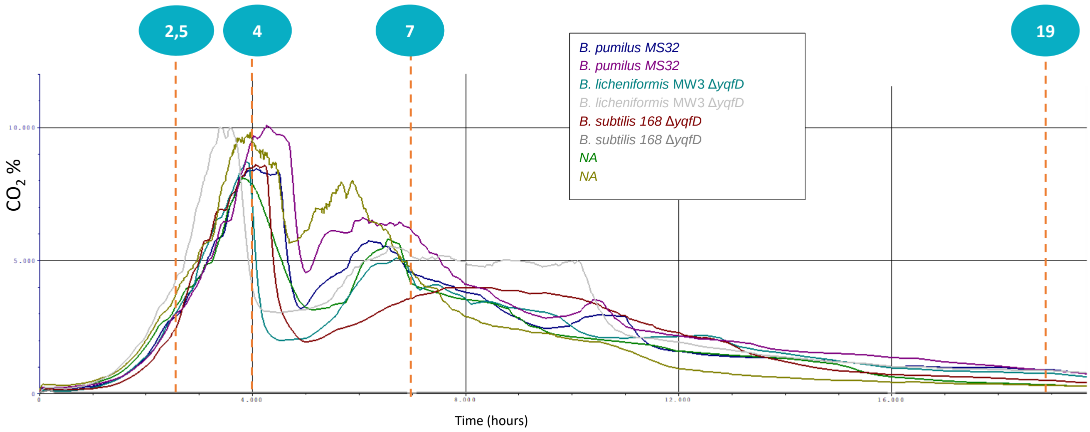


FIGURE 3.8: Carbon dioxide measurements for the fermentation runs of *B. pumilus* MS32 and *B. licheniformis* MW3 Δ *yqfD*, *B. subtilis* 168 is also included. Sampling points are depicted.

3.4 RNA-seq Analysis

In this section, the results of processing the *B. pumilus* and *B. licheniformis* libraries is described, with special focus on identification of small regulatory RNAs and expression profiles revealing the culture behavior of these *Bacillus* species during fermentation conditions. These results complement and expand the comparative genomic analysis presented in the previous section by adding the layer of transcriptional activity across time points. This approach highlights species-specific differences in regulation and transcription of genomic features which are relevant to consider for selection, development, and optimization of industrial production hosts.

3.4.1 Quality assessment of RNA-seq libraries

In the manuscript "RNA of high yield, integrity and purity from industrial *Bacillus*, an improved method" (Chapter 6) it was shown that the optimized RNA isolation protocol successfully preserved the integrity of the RNA samples, which was reflected in the reported TIN and RIN values. Table 3.11 presents a more detailed summary of quality metrics for each of the RNA-libraries generated for *B. pumilus* MS32 and *B. licheniformis* MW3 Δ *yqfD* at each sampling point.

TABLE 3.11: Quality metrics for the RNA-seq raw data generated from *B. pumilus* MS32 and *B. licheniformis* MW3 Δ *yqfD* fermentation samples take at different time points after bioreactor inoculation. Analysis was done with FastP [52], SortmeRNA [181], and RSEQC [335] software packages.

Time (h)	Sample ID	Total Reads	%passQC	%rRNA	TIN mean	TIN median
B. pumilus MS32						
2.5	SDIA_t_78	6,606,758	99.6	2.42	87.9	92.3
2.5	SDIA_t_82	5,935,764	99.6	0.24	87.5	92.0
2.5	SDIA_t_165	6,774,520	99.6	0.13	88.1	92.7
4	SDIA_t_87	5,901,122	99.6	2.20	88.1	92.1
4	SDIA_t_96	7,011,948	99.7	0.45	88.4	92.3
4	SDIA_t_118	7,618,086	99.6	0.14	88.3	92.6
7	SDIA_t_134	7,454,614	99.3	0.13	87.5	91.4
7	SDIA_t_135	7,097,390	99.2	0.32	87.4	91.0
7	SDIA_t_124	5,440,014	99.1	0.63	85.0	89.9
19	SDIA_t_166	6,522,936	99.5	63.06	87.0	91.6
19	SDIA_t_137	6,795,884	99.6	52.90	87.6	91.9
19	SDIA_t_139	5,824,424	99.2	0.35	88.2	92.0
B. licheniformis DSM13						
2.5	SDIA_t_80	8,587,532	99.3	0.32	87.3	92.4
2.5	SDIA_t_84	9,914,484	99.5	43.81	86.6	92.1
2.5	SDIA_t_85	10,963,976	99.4	1.79	87.9	93.0
4	SDIA_t_89	9,300,750	99.4	0.84	85.4	91.5
4	SDIA_t_133	6,216,532	99.4	18.01	85.5	91.5
4	SDIA_t_120	5,992,058	99.5	0.46	84.9	91.1
7	SDIA_t_70	6,408,938	99.6	7.51	87.2	91.9
7	SDIA_t_122	6,833,640	99.1	58.99	85.3	90.9
7	SDIA_t_136	5,958,576	99.1	1.72	86.4	91.4
19	SDIA_t_18	6,780,438	99.7	62.44	85.0	90.5
19	SDIA_t_138	6,723,646	99.1	48.10	86.8	91.9
19	SDIA_t_76	7,955,478	99.0	70.57	77.2	82.6

In all cases more than 99% of the reads passed the QC filtering and all libraries contained enough reads for further analysis. Regarding rRNAs, despite the use of custom made probes for the rRNA depletion step of the cDNA library preparation, some samples still had a high percentage of rRNA. Particularly, in the late sampling points of *B. licheniformis*. The amount of reads remaining still allowed further analysis.

Table 3.12 presents a summary of the sequencing data after read alignment to the corresponding *B. pumilus* and *B. licheniformis* genomes, this is part of the report generated by the READemption align subcommand [104].

TABLE 3.12: Read alignment summary of the RNA-seq data generated for *B. pumilus* MS32 and *B. licheniformis* MW3 Δ *yqfD*.

Timepoint	Sample ID	Input reads	Aligned	Unaligned	Uniquely aligned	Alignments	Aligned %	Uniquely aligned %
B. pumilus MS32								
2.5	SDIA_t_78	6,606,758	6,548,925	56,169	6,530,035	6,582,640	99.12	99.71
2.5	SDIA_t_82	5,935,764	5,888,079	45,751	5,870,626	5,919,385	99.2	99.7
2.5	SDIA_t_165	6,774,520	6,701,278	69,470	6,681,054	6,738,037	98.92	99.7
4	SDIA_t_87	5,901,122	5,837,415	60,265	5,818,779	5,869,590	98.92	99.68
4	SDIA_t_96	7,011,948	6,938,808	72,286	6,914,705	6,980,105	98.96	99.65
4	SDIA_t_118	7,618,086	7,548,628	65,480	7,522,816	7,595,640	99.09	99.66
7	SDIA_t_134	7,454,614	7,334,154	111,872	7,312,554	7,370,434	98.38	99.71
7	SDIA_t_135	7,097,390	6,977,483	111,591	6,958,912	7,009,624	98.31	99.73
7	SDIA_t_124	5,440,014	5,305,604	121,926	5,288,890	5,334,099	97.53	99.68
19	SDIA_t_166	6,522,936	6,399,840	123,032	6,373,617	6,463,346	98.11	99.59
19	SDIA_t_137	6,795,884	6,628,930	166,920	6,586,873	6,738,005	97.54	99.37
19	SDIA_t_139	5,824,424	5,733,384	78,286	5,719,877	5,755,013	98.44	99.76
B. licheniformis MW3								
2.5	SDIA_t_80	8,587,532	8,424,032	158,894	8,394,365	8,862,414	98.1	99.65
2.5	SDIA_t_84	9,914,484	9,726,660	171,092	5,467,856	78,951,502	98.11	56.22
2.5	SDIA_t_85	10,963,976	10,737,344	222,412	10,539,685	13,865,376	97.93	98.16
4	SDIA_t_89	9,300,750	8,718,248	563,732	8,639,041	10,032,016	93.74	99.09
4	SDIA_t_133	6,216,532	6,023,000	181,858	4,949,362	24,274,770	96.89	82.17
4	SDIA_t_120	5,992,058	5,907,308	75,638	5,879,220	6,359,804	98.59	99.52
7	SDIA_t_70	6,408,938	6,309,320	96,222	5,846,181	13,932,781	98.45	92.66
7	SDIA_t_122	6,833,640	6,703,289	112,703	2,768,585	74,322,438	98.09	41.3
7	SDIA_t_136	5,958,576	5,873,602	83,192	5,770,755	7,555,333	98.57	98.25
19	SDIA_t_18	6,780,438	6,724,033	56,369	2,543,034	77,666,281	99.17	37.82
19	SDIA_t_138	6,723,646	6,646,553	60,027	3,492,817	60,643,258	98.85	52.55
19	SDIA_t_76	7,955,478	6,085,970	1,810,160	1,061,989	80,282,979	76.5	17.45

For most of the libraries more than 97% of the input reads mapped to the reference genomes, which indicates that the RNA-seq data was suitable for further transcriptomic evaluation. Libraries with an inferior percentage of uniquely aligned reads correspond to libraries in which rRNA derived reads were most abundant.

3.4.2 Predicted RNAs

The generated transcriptomic data was used to predict candidate active sRNAs using the Annogesic [363] software suite. Table 3.13 presents a summary of the candidate sRNAs found in *B. pumilus* and *B. licheniformis* MW3 Δ *yqfD*, below each total, other descriptors of the predicted sRNAs are given.

TABLE 3.13: Candidate sRNAs in *B. pumilus* MS32 and *B. licheniformis* MW3 Δ *yqfD* genomes. Below each total prediction a breakdown of further characteristics is depicted.

	B. pumilus			***B. licheniformis***		
	Antisense	Intergenic	Total	Antisense	Intergenic	Total
sRNA candidates	30	13	43	41	35	76
Normalized free energy change of the secondary structure is below -0.05	30	13	43	41	35	76
Ends with terminator	1	5	6	1	18	19
No conflict with sORFs	19	6	25	28	24	52
Homology to sRNA database	5	5	10	4	7	11

Additionally, the sRNA prediction by Annogesic [363] was complemented by the analysis against the covariance model collection of the RFAM database [121]. The analysis identified further non-conding RNAs within the *Bacillus* genomes. Table 3.14 presents a summary of the matches obtained for *B. pumilus* MS32 and *B. licheniformis* MW3, the results of *B. subtilis* 168 are included for broader comparison purposes. The corresponding RFAM covariance model and Clan identifiers are depicted.

TABLE 3.14: Identified RNAs in *B. pumilus* MS32, *B. licheniformis* MW3 and *B. subtilis* 168 using Infernal [238] against the RFAM database [121].

Description	RFAM	CLAN	*B. pumilus*	*B. licheniformis*	*B. subtilis*
5S ribosomal RNA	RF00001	CL00113	8	7	10
5′ ureB small RNA	RF02514	-	0	0	1
6S / SsrS RNA	RF00013	-	2	2	2
Bacillaceae-1 RNA	RF01690	-	0	23	9
Bacillus asRNA 0872	RF02662	-	0	1	0
Bacillus-plasmid RNA	RF01691	-	1	0	1
Bacillus SR6 antitoxin	RF02892	-	0	0	1
Bacillus sRNA ncr1015	RF02449	-	0	1	1
Bacillus sRNA ncr1175	RF02450	-	0	0	1
Bacillus sRNA ncr1241	RF02451	-	0	1	1
Bacillus sRNA ncr1575	RF02452	-	0	0	1
Bacillus sRNA ncr952	RF02453	-	0	0	1
Bacillus sRNA ncr982	RF02454	-	1	0	1
Bacillus tryptophan operon leader	RF02370	-	1	1	1
Bacterial large signal recognition particle RNA	RF01854	CL00003	1	1	1
Bacterial large subunit ribosomal RNA	RF02541	CL00112	8	7	10
Bacterial RNase P class B	RF00011	CL00002	1	1	1
Bacterial small subunit ribosomal RNA	RF00177	CL00111	8	7	10
BsrC	RF01410	-	1	1	2
BsrF	RF01411	-	1	0	1
BsrG	RF01412	-	1	3	3
Cobalamin riboswitch	RF00174	CL00101	1	1	1
cspA thermoregulator	RF01766	-	3	1	1
Cyclic di-GMP-I riboswitch	RF01051	CL00126	1	1	0
DicF RNA	RF00039	-	0	0	1
DUF3800-VIII RNA	RF03075	-	0	0	1
epsC RNA	RF01735	-	1	1	1
FMN riboswitch (RFN element)	RF00050	-	2	2	2
FsrA	RF02273	-	1	1	1
glmS glucosamine-6-phosphate activated ribozyme	RF00234	-	1	1	1
Glycine riboswitch	RF00504	CL00125	2	1	1
Group I catalytic intron	RF00028	-	0	0	1
Guanidine-I riboswitch	RF00442	-	1	2	2
JUMPstart RNA	RF01707	-	0	1	0
L31-Firmicutes ribosomal protein leader	RF03156	CL00118	1	1	0
Listeria snRNA rli23	RF01458	-	19	0	0
Lysine riboswitch	RF00168	-	2	2	2
M-box riboswitch (ykoK leader)	RF00380	-	1	1	1
pan motif	RF01749	-	1	1	1
PreQ1 riboswitch	RF00522	-	1	1	1
Purine riboswitch	RF00167	CL00123	4	4	5
PyrG leader	RF02371	-	1	1	1
PyrR binding site	RF00515	-	3	3	3
Ribosomal protein L10 leader	RF00557	-	1	1	1
Ribosomal protein L13 leader	RF00555	-	1	1	1
Ribosomal protein L19 leader	RF00556	-	1	1	1
Ribosomal protein L20 leader	RF00558	-	1	1	1
Ribosomal protein L21 leader	RF00559	-	1	1	1
RNA Staph. aureus E (RoxS)	RF01820	-	1	1	1
S10-Clostridia ribosomal protein leader	RF03136	-	1	1	1
SAM riboswitch (S box leader)	RF00162	CL00012	9	10	11
SR1 sRNA	RF02376	-	1	1	1
SurA sRNA	RF02377	-	0	1	1
SurC sRNA	RF02378	-	0	0	1
T-box leader	RF00230	-	15	14	15
TPP riboswitch (THI element)	RF00059	-	4	4	5
transfer-messenger RNA	RF00023	CL00001	1	1	1
tRNA	RF00005	CL00001	81	72	86
ydaO/yuaA leader	RF00379	-	2	2	2
yjdF RNA	RF01764	-	0	0	1
ylbH leader	RF00516	-	1	1	1
yybP-ykoY manganese riboswitch	RF00080	-	1	0	0

Annogesic [363] found for *B. pumilus* MS32 a total of 43 potential sRNAs, with 30 putative antisense sRNAs and 13 intergenic, their size ranged from 60 to 493 nucleotides. More sRNAs were predicted for *B. licheniformis* MW3 Δ *yqfD*, with a total of 76; 41 antisense, 35 intergenic, and a size range of 60 to 451 nt.

3.4.3 Antisense activity of predicted sRNAs

The small RNAs predicted to be antisense by Annogesic were further investigated. Tables 3.15 and 3.16 depicts the coordinates, strand, size, and feature encoded opposite to these candidate sRNAs. A Pearson correlation coefficient and the corresponding p-value between the transcriptional trajectory of the putative sRNA and the antisense feature are also presented. The maximum TPM activity of the sRNA (triplicate mean) and the sampling time point at which it occurs are included.

TABLE 3.15: Transcriptional activity of predicted antisense RNA in *B. pumilus* and their putative target predicted by Annogesic [363]. Pearson correlations and the corresponding p-values are presented.

sRNA	Start	Stop	Strand	Size (nt)	Antisense feature	Correlation Coefficient	p-val	Max. mean TPM of asRNA	Max. asRNA TPM at:
srna29	3100710	3101162	+	452	hypothetical protein	-0.999	0.001	132,054	7h
srna22	2849449	2849865	+	416	RapH phosphatase inhibitor	-0.944	0.056	102,961	2.5h
srna32	3190037	3190520	+	483	C40 family peptidase	-0.938	0.062	247,79	7h
srna33	3190640	3190700	+	60	C40 family peptidase	-0.917	0.083	174,816	19
srna18	2606677	2607164	-	487	YslB family protein	-0.844	0.156	91,804	7h
srna35	3360207	3360390	-	183	YwpF-like family protein	-0.768	0.232	103,12	2.5h
srna42	3734331	3734564	-	233	PhzF family phenazine biosynthesis protein	-0.767	0.233	99,007	4h
srna5	757339	757649	+	310	DUF1540 domain-containing protein	-0.754	0.246	282,826	7h
srna0	51710	52201	-	491	DUF348 domain-containing protein	-0.743	0.257	1650,3	7h
srna25	3001335	3001471	-	136	YhcN/YlaJ family sporulation lipoprotein	-0.682	0.318	75,355	2.5h
srna26	3001460	3001609	-	149	YhcN/YlaJ family sporulation lipoprotein	-0.681	0.319	63,526	2.5h
srna15	2365868	2365983	+	115	RNA polymerase sigma factor RpoD	-0.625	0.375	151,073	19
srna27	3001671	3001902	-	231	YhcN/YlaJ family sporulation lipoprotein	-0.542	0.458	69,443	2.5h
srna7	936946	937271	-	325	ABC transporter ATP-binding protein	-0.504	0.496	317,434	4h
srna34	3359923	3360190	-	267	YwpF-like family protein	-0.452	0.548	70,211	2.5h
srna24	2924412	2924523	-	111	sensor histidine kinase	-0.426	0.574	131,19	7h
srna23	2918412	2918746	-	334	DUF378 domain-containing protein	-0.384	0.616	76,61	2.5h
srna13	1961420	1961759	+	339	LysM peptidoglycan-binding domain-containing protein	-0.048	0.952	951,642	7h
srna20	2754924	2755233	-	309	NCS2 family permease	0.032	0.968	136,494	4h
srna11	1134759	1135179	+	420	DegV family protein	0.084	0.916	108,669	19
srna40	3695673	3695987	+	314	CDP-glycerol glycerophosphotransferase family protein	0.136	0.864	95,282	4h
srna41	3696165	3696252	+	87	CDP-glycerol glycerophosphotransferase family protein	0.15	0.85	137,589	4h
srna16	2377666	2377943	+	277	HD family phosphohydrolase	0.207	0.793	217,115	19
srna14	2092474	2092676	+	202	helix-turn-helix domain-containing protein	0.274	0.726	78,547	2.5h
srna4	604196	604689	+	493	ABC-F type ribosomal protection protein	0.382	0.618	71,394	4h
srna21	2838318	2838661	-	343	energy-coupled thiamine transporter ThiT	0.457	0.543	77,637	4h
srna37	3548822	3549079	-	257	cation acetate symporter	0.765	0.235	281,885	19
srna19	2701447	2701510	-	63	ABC transporter permease	0.969	0.031	174,109	19

TABLE 3.16: Transcriptional activity of predicted antisense RNA in *B. licheniformis* and their putative target predicted by Annogesic [363]. Pearson correlations and the corresponding p-values are presented.

sRNA	Start	Stop	Strand	Size	Antisense feature	Correlation Coefficient	pval	Max. mean TPM of asRNA	Max. asRNA TPM at:
srna59	3183471	3183726	-	255	pyridoxal phosphate-dependent aminotransferase	-0.981	0.019	199,175	19
srna72	3713716	3713982	-	266	GNAT family N-acetyltransferase	-0.925	0.075	83,475	4h
srna75	3805080	3805267	-	187	penicillin-binding protein	-0.87	0.13	164,207	4h
srna56	3091258	3091370	-	112	membrane protein insertion efficiency factor YidD	-0.84	0.16	162,04	2.5h
srna55	3071722	3072123	-	401	glutamate-aspartate/proton symporter GltP	-0.83	0.17	81,663	4h
srna74	3804766	3805089	-	323	penicillin-binding protein	-0.789	0.211	125,686	4h
srna71	3696821	3696941	+	120	large conductance mechanosensitive channel protein MscL	-0.502	0.498	266,685	4h
srna54	3034347	3034798	-	451	NCS2 family permease	-0.188	0.812	293,473	4h
srna53	3033997	3034337	-	340	NCS2 family permease	-0.145	0.855	406,789	4h
srna39	2282753	2283015	-	262	cold-shock protein CspD	-0.09	0.91	114,657	2.5h
srna19	1226941	1227095	-	154	YjzC family protein	-0.038	0.962	127,857	2.5h
srna16	1053164	1053255	+	91	NAD(P)H-binding protein	0.028	0.972	176,367	19
srna46	2551137	2551398	+	261	M73 family metallopeptidase	0.523	0.477	84,342	2.5h
srna62	3377496	3377872	-	376	GrpB family protein	0.55	0.45	170,641	19
srna31	1652795	1653104	+	309	YhcN/YlaJ family sporulation lipoprotein	0.587	0.413	311,534	19
srna45	2551055	2551132	+	77	M73 family metallopeptidase	0.596	0.404	91,22	2.5h
srna26	1564663	1564879	-	216	phosphoenolpyruvate–protein phosphotransferase	0.735	0.265	105,303	2.5h
srna68	3491361	3491450	+	89	type I glyceraldehyde-3-phosphate dehydrogenase	0.742	0.258	114,426	2.5h
srna69	3491457	3491691	+	234	type I glyceraldehyde-3-phosphate dehydrogenase	0.746	0.254	114,607	2.5h
srna24	1424740	1425128	+	388	hypothetical protein	0.775	0.225	128,884	2.5h
srna47	2627472	2627877	+	405	GatB/YqeY domain-containing protein	0.781	0.219	124,447	2.5h
srna4	129385	129817	-	432	30S ribosomal protein S7	0.796	0.204	210,547	2.5h
srna18	1108980	1109129	+	149	S8 family peptidase	0.802	0.198	678,006	19
srna67	3490909	3491191	+	282	phosphoglycerate kinase	0.813	0.187	129,232	2.5h
srna64	3486459	3486842	+	383	phosphopyruvate hydratase	0.822	0.178	130,04	2.5h
srna42	2356213	2356479	+	266	non-specific DNA-binding protein Hbs	0.844	0.156	226,953	2.5h
srna73	3770580	3770968	+	388	DegT/DnrJ/EryC1/StrS family aminotransferase	0.855	0.145	144,419	19
srna70	3491719	3492098	+	379	type I glyceraldehyde-3-phosphate dehydrogenase	0.875	0.125	174,577	2.5h
srna65	3488086	3488330	+	244	2,3-bisphosphoglycerate-independent phosphoglycerate mutase	0.887	0.113	117,562	2.5h
srna3	118278	118396	-	118	50S ribosomal protein L1	0.899	0.101	232,257	2.5h
srna32	1781368	1781603	-	235	50S ribosomal protein L19	0.915	0.085	173,73	2.5h
srna66	3489363	3489592	+	229	triose-phosphate isomerase	0.922	0.078	117,741	2.5h
srna48	2643960	2644222	-	262	30S ribosomal protein S20	0.937	0.063	292,637	2.5h
srna6	137955	138087	-	132	50S ribosomal protein L29	0.949	0.051	361,228	2.5h
srna30	1635034	1635339	-	305	2-oxo acid dehydrogenase subunit E2	0.961	0.039	178,185	2.5h
srna28	1634215	1634541	-	326	pyruvate dehydrogenase complex E1 component subunit beta	0.973	0.027	109,867	2.5h
srna29	1634707	1635000	-	293	pyruvate dehydrogenase complex E1 component subunit beta	0.977	0.023	112,015	2.5h
srna5	131295	131609	-	314	elongation factor G	0.999	0.001	150,864	2.5h

A total of 30 antisense RNAs (asRNAs) were predicted by Annogesic [363] for *B. pumilus* MS32 and 41 for *B. licheniformis* MW3. Table 3.15 shows that the putative asRNAs with higher transcriptional activities in *B. pumilus* MS32 were srna0 and srna13. These candidates presented homology to Bpsr193 and Bpsr92, respectively (previously identified sRNAs in *B. altitudinis* SC11 [353]). Bpsr193/srna0 matched the transcriptional trajectory depicted by cluster 17 (Figure 3.11), with a more than 4-fold increased activity between 2.5 hours and the transition point, followed by TPM values of 1650,3 and 1473,5 at 7 and 19 hours, respectively. The antisense gene for this candidate sRNA is an uncharacterized protein presenting a DUF348 domain, and there was a negative correlation indicated between these features (Table 3.15). The candidate Bpsr92/srna13 was also found within the cluster 17 (Figure 3.11) and it is encoded in the opposite strand of a LysM peptidoglycan-binding domain-containing protein. In *B. altitudinis* SC11 Bpsr92 is 343 nt in length [353] while the putative srna13 in *B. pumilus* MS32 is 339 nt, no further characterization was reported for this sRNA.

In *B. licheniformis* two candidate asRNAs (srna54 and srna53) are encoded in the opposite strand of a protein annotated as "NCS2 family permease" (Table 3.16). According to the Proteinortho [188] analysis, this protein corresponds to hypoxanthineguanine permease PbuO in *B. subtilis* and it is also present in *B. pumilus* MS32. Both candidate asRNAs reached maximum transcriptional activity at 4 hours and presented a strong 11-fold induction with respect to the 2.5 hour sampling point, transcription was afterwards down-regulated to initial values during the stationary phase. This transcriptional trajectory is depicted in cluster 15 (Figure 3.13). Table 3.15 shows that the putative srna20 of MS32 was also located antisense of *pbuO*, the clustering analysis assigned this asRNA to cluster 16 (Figure 3.10), which follows a similar trajectory to that of *B. licheniformis*, with the difference of not being so strongly induced relative to 2.5 hours (1.6 fold increase in activity). From Table 3.16 it is shown that the asRNA with highest transcriptional activity in *B. licheniformis* was the candidate srna18, which corresponds to the previously characterized AprAs asRNA which regulates the expression of the Apr protease [139].

3.4.4 Determination and clustering of transcriptional profiles across fermentations

The transcriptional activities for *B. pumilus* MS32 and *B. licheniformis* MW3 have been summarized in a catalog of different transcriptional profiles across the fermentation sampling points. Figures 3.9, 3.10 and 3.11 represent the clusters determined for *B. pumilus* by DP_GP_cluster[220]. For *B. licheniformis* the corresponding clusters are presented in figures 3.12, 3.13, 3.14 and 3.15.

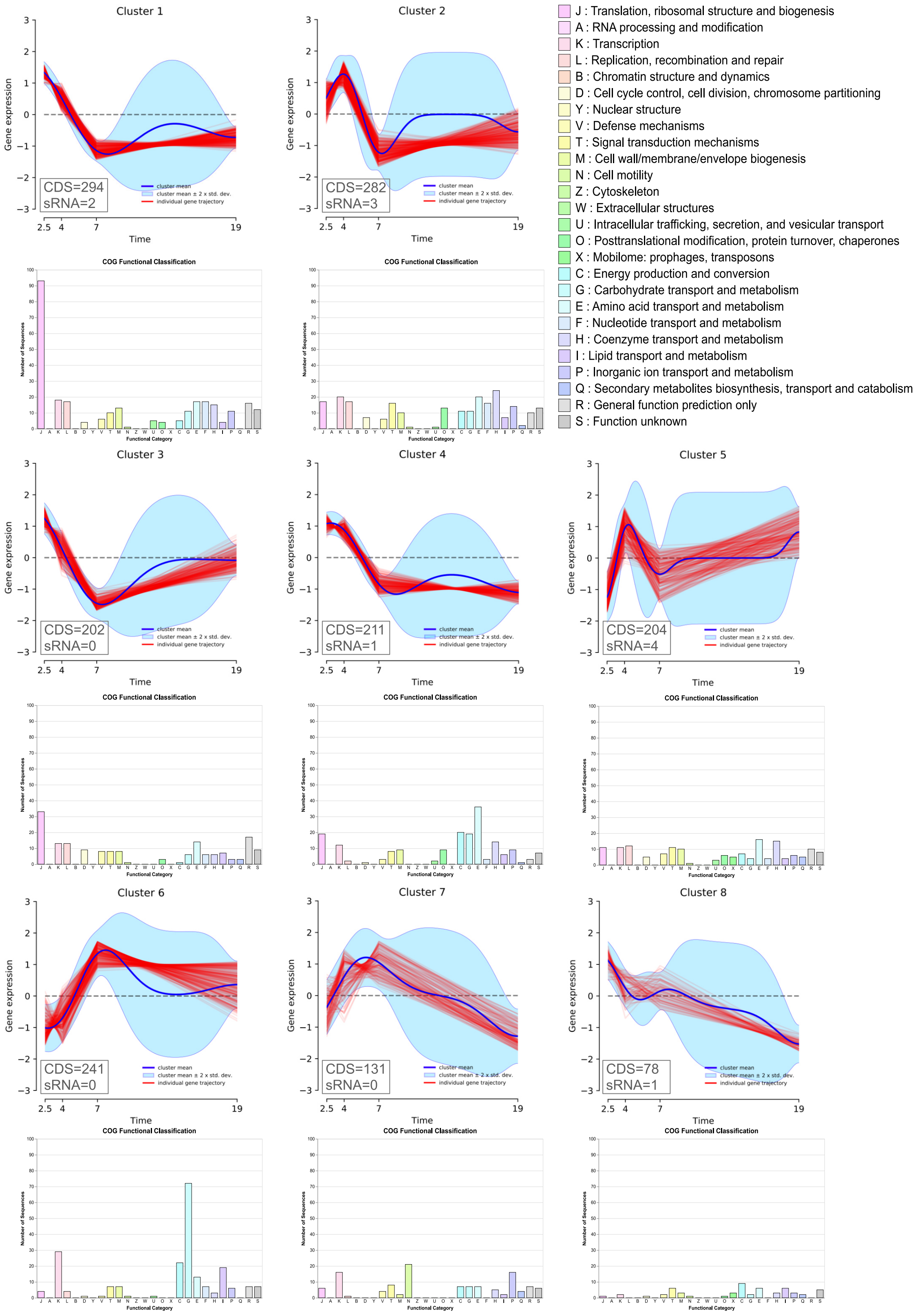


FIGURE 3.9: Transcriptional activities of *B. pumilus* MS32 clustered into profiles by DPGP analysis. Total CDS and sRNA features within each cluster are depicted in a corner box. Distribution of COG functional assignments is shown below every cluster. Blue lines= cluster mean, red lines = individual trajectories and light blue = cluster mean ± 2 x std. deviation.

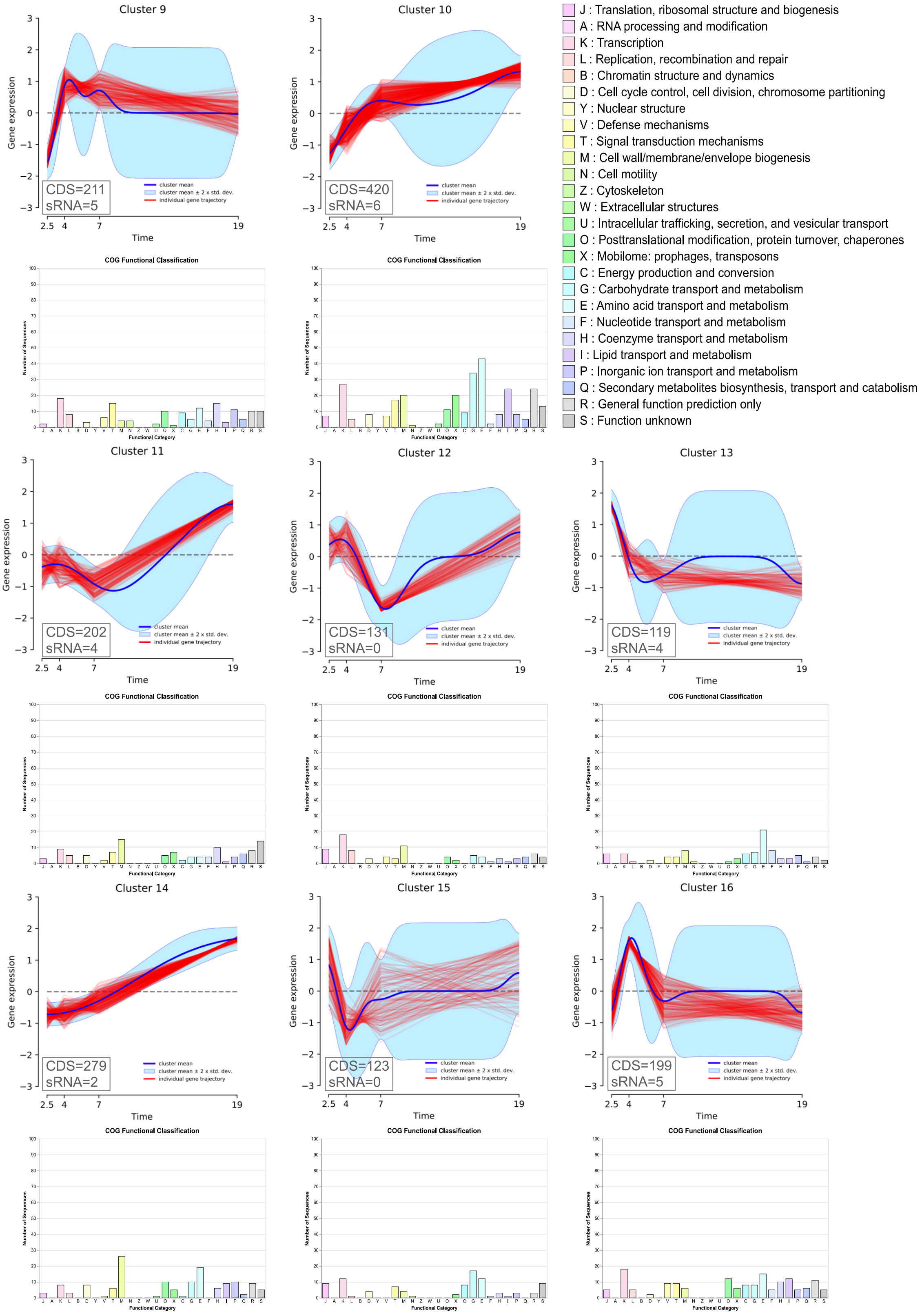


FIGURE 3.10: Transcriptional activities of *B. pumilus* MS32 clustered into profiles by DPGP analysis. Total CDS and sRNA features within each cluster are depicted in a corner box. Distribution of COG functional assignments is shown below every cluster. Blue lines= cluster mean, red lines = individual trajectories and light blue = cluster mean ± 2 x std. deviation.

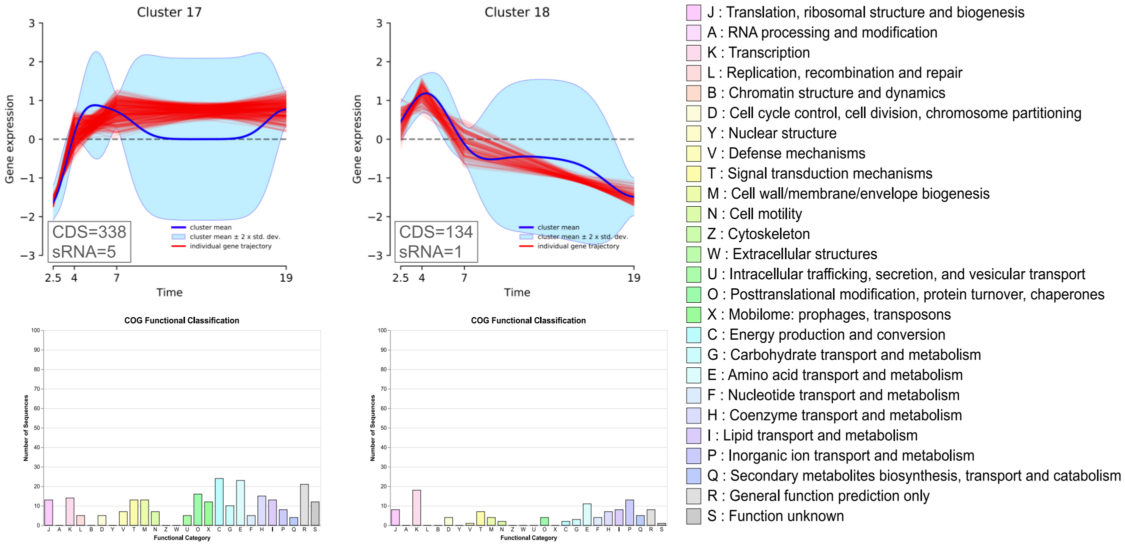


FIGURE 3.11: Transcriptional activities of *B. pumilus* MS32 clustered into profiles by DPGP analysis. Total CDS and sRNA features within each cluster are depicted in a corner box. Distribution of COG functional assignments is shown below every cluster. Blue lines= cluster mean, red lines = individual trajectories and light blue = cluster mean ± 2 x std. deviation.

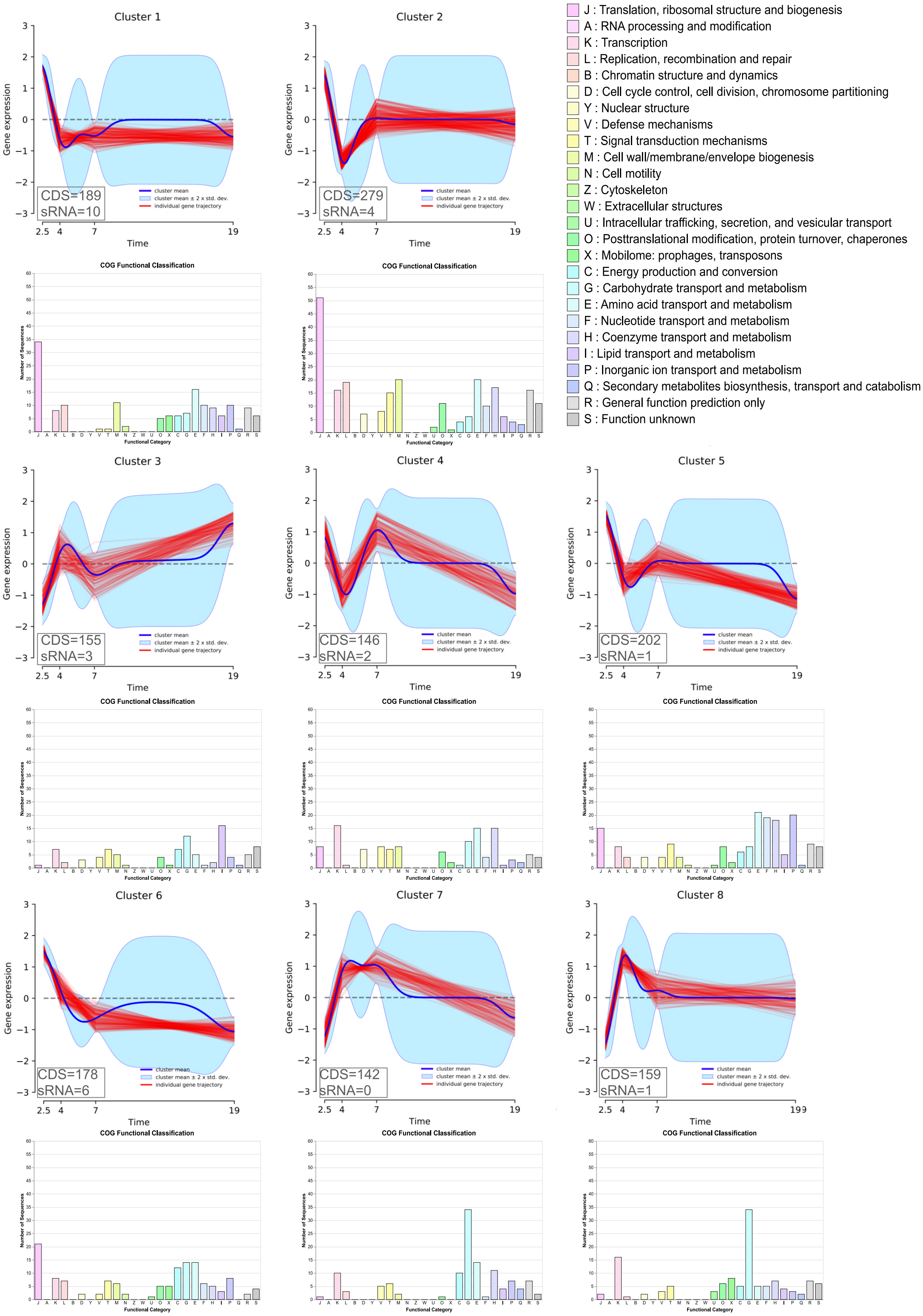


FIGURE 3.12: Transcriptional activities of *B. licheniformis* MW3 Δ *yqfD* clustered into profiles by DPGP analysis. Total CDS and sRNA features within each cluster are depicted in a corner box. Distribution of COG functional assignments is shown below every cluster. Blue lines= cluster mean, red lines = individual trajectories and light blue = cluster mean ± 2 x std. deviation.

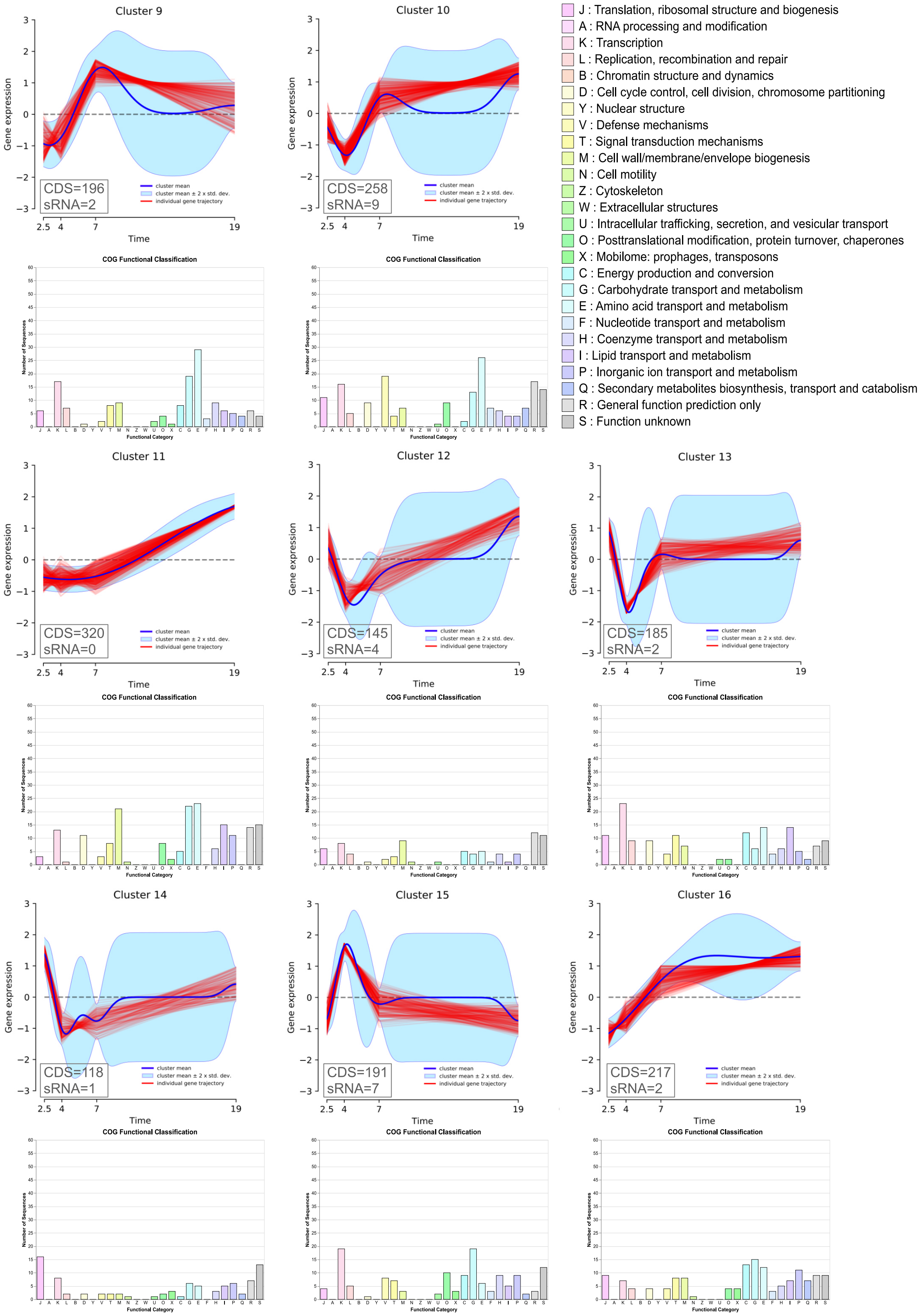


FIGURE 3.13: Transcriptional activities of *B. licheniformis* MW3 Δ *yqfD* clustered into profiles by DPGP analysis. Total CDS and sRNA features within each cluster are depicted in a corner box. Distribution of COG functional assignments is shown below every cluster. Blue lines= cluster mean, red lines = individual trajectories and light blue = cluster mean ± 2 x std. deviation.

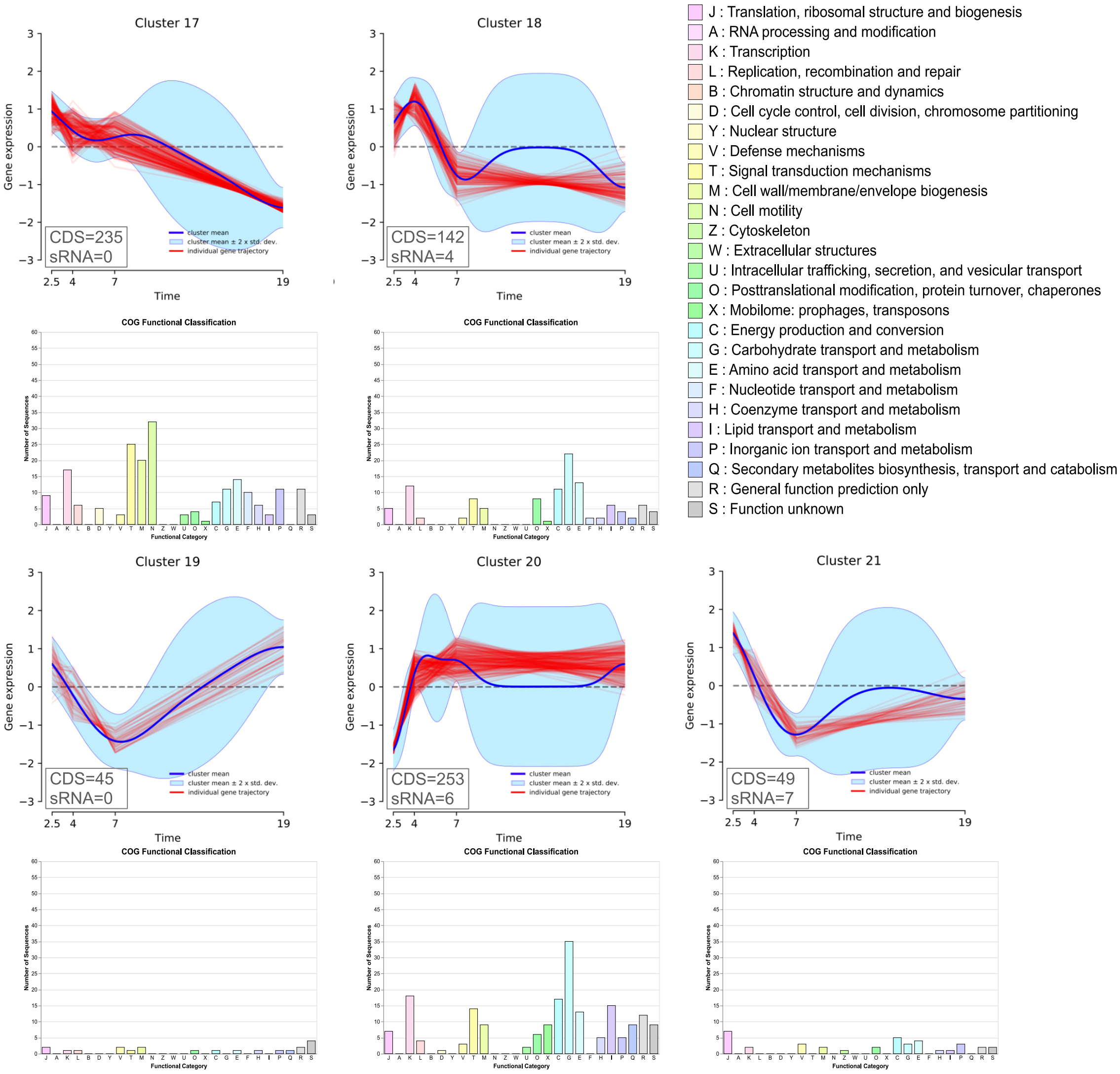


FIGURE 3.14: Transcriptional activities of *B. licheniformis* MW3 Δ *yqfD* clustered into profiles by DPGP analysis. Total CDS and sRNA features within each cluster are depicted in a corner box. Distribution of COG functional assignments is shown below every cluster. Blue lines= cluster mean, red lines = individual trajectories and light blue = cluster mean ± 2 x std. deviation.

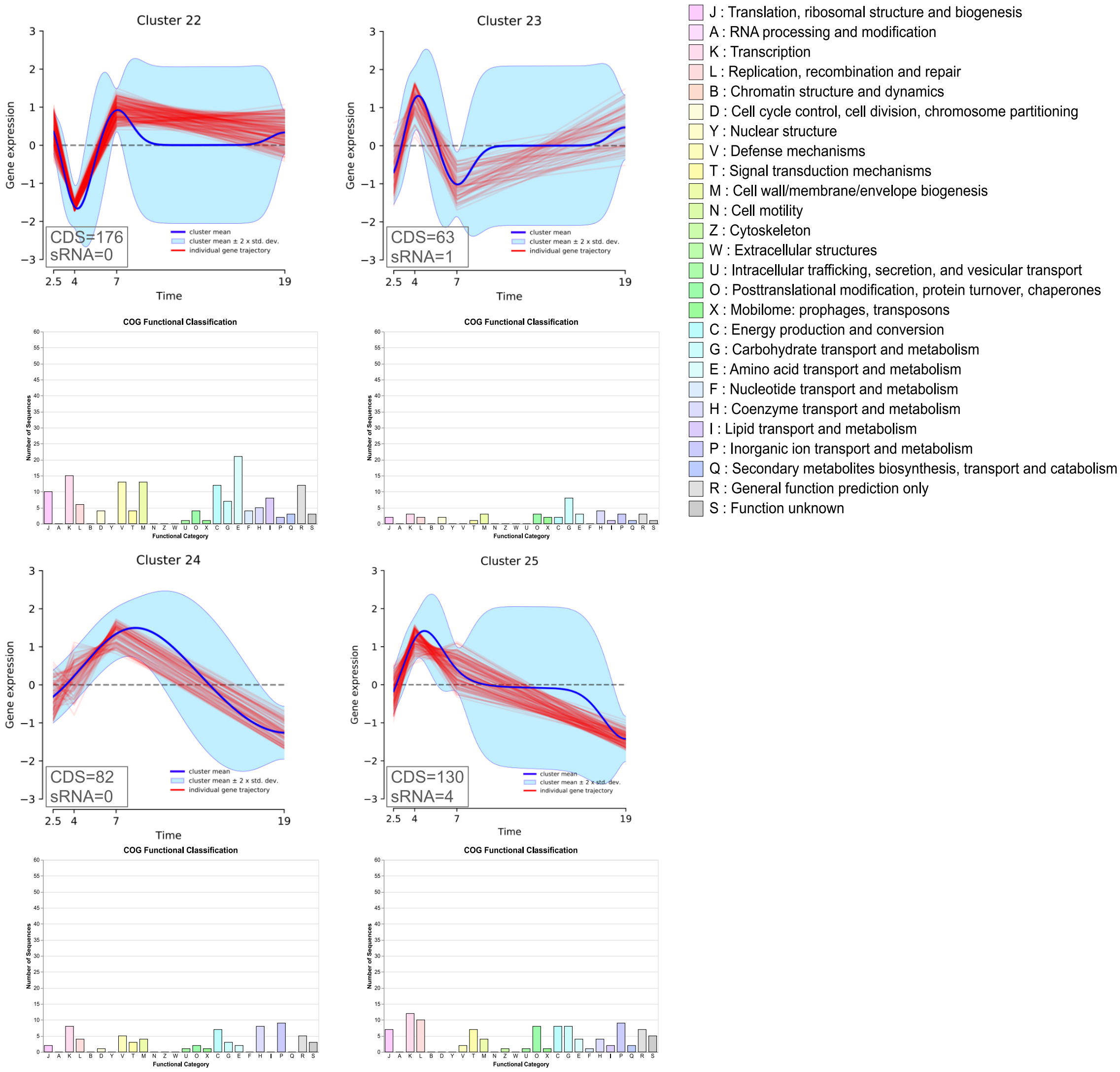


FIGURE 3.15: Transcriptional activities of *B. licheniformis* MW3 Δ *yqfD* clustered into profiles by DPGP analysis. Total CDS and sRNA features within each cluster are depicted in a corner box. Distribution of COG functional assignments is shown below every cluster. Blue lines= cluster mean, red lines = individual trajectories and light blue = cluster mean ± 2 x std. deviation.

The DP_GP_cluster[220] analysis determined 18 clusters of features with similar transcriptional profiles for *B. pumilus* MS32 and 25 for *B. licheniformis* MW3 Δ *yqfD*. This could be related to *B. licheniformis* having a bigger genome encoding more proteins than *B. pumilus* (4164 and 3712 ORFs, respectively, Table 3.1).

For *B. pumilus* MS32 each cluster contained in average 213 features, with clusters 10 and 8 showing the maximum (426) and minimum (79), respectively. Cluster 10 also presented the most predicted sRNAs, while clusters 6, 3, 7, 12, and 15 had no sRNA following the corresponding trajectories. For cluster 10, 156 genes could be associated with a GO term, being "Cell Differentiation" the most abundant Biological Process (16) and "Hydrolase activity" the most abundant Molecular Function (24). COG assignment revealed that categories: "Amino acid transport and metabolism", "Carbohydrate transport and metabolism", and "Transcription" were the categories with the highest counts, in that order. Further inspection of cluster elements showed that key elements such as *secE*, *sigH*, *spo0A*, *abbA*, *dppA* and *aprE* followed the transcriptional trajectory described by cluster 10 members. Which started with low transcription at 2.5h gradually increasing at 4 and 7 hours to finally peaked at the final sampling point.

In the case of *B. licheniformis*, the average number of members per cluster was 173.24. Cluster 11 grouped the most proteins (320) while cluster 19 had the fewest (45), no predicted sRNA was found within these clusters. Cluster 1 presented the most predicted sRNAs (10) and the major Biological Process associated with it was "Cellular amino acid metabolic process". Similarly to cluster 10 of *B. pumilus*, the most abundant Biological Process and Molecular Function categories in cluster 11 were "Cell Differentiation" (38) and "Hydrolase activity" (25). The top three COGs in this cluster were: "Amino acid transport and metabolism", "Carbohydrate transport and metabolism" and "Cell wall/membrane/envelope biogenesis". Unlike in *B. pumilus*, the transcriptomic trajectory of cluster 11 is rather stable and below average at the first three time points and peaks at 19 hours. Members of cluster 11 include: *degQ*, *aprE*, and *cspC*.

3.4.5 Diferentially Expressed Genes

From the quantification step READemption [104] generated raw counts, which were then processed by DESeq package for Differential Gene Expression Analysis. The analysis is useful to contrast between the sampling points and revealed which genes are significantly up and down regulated during the different phases of the fermentation. These genes can be linked to cell status and their response to the fermentation conditions. Table 3.16 depicts volcano plots of *B. pumilus* MS32 and *B. licheniformis* MW3 Δ *yqfD* for every time point comparison.

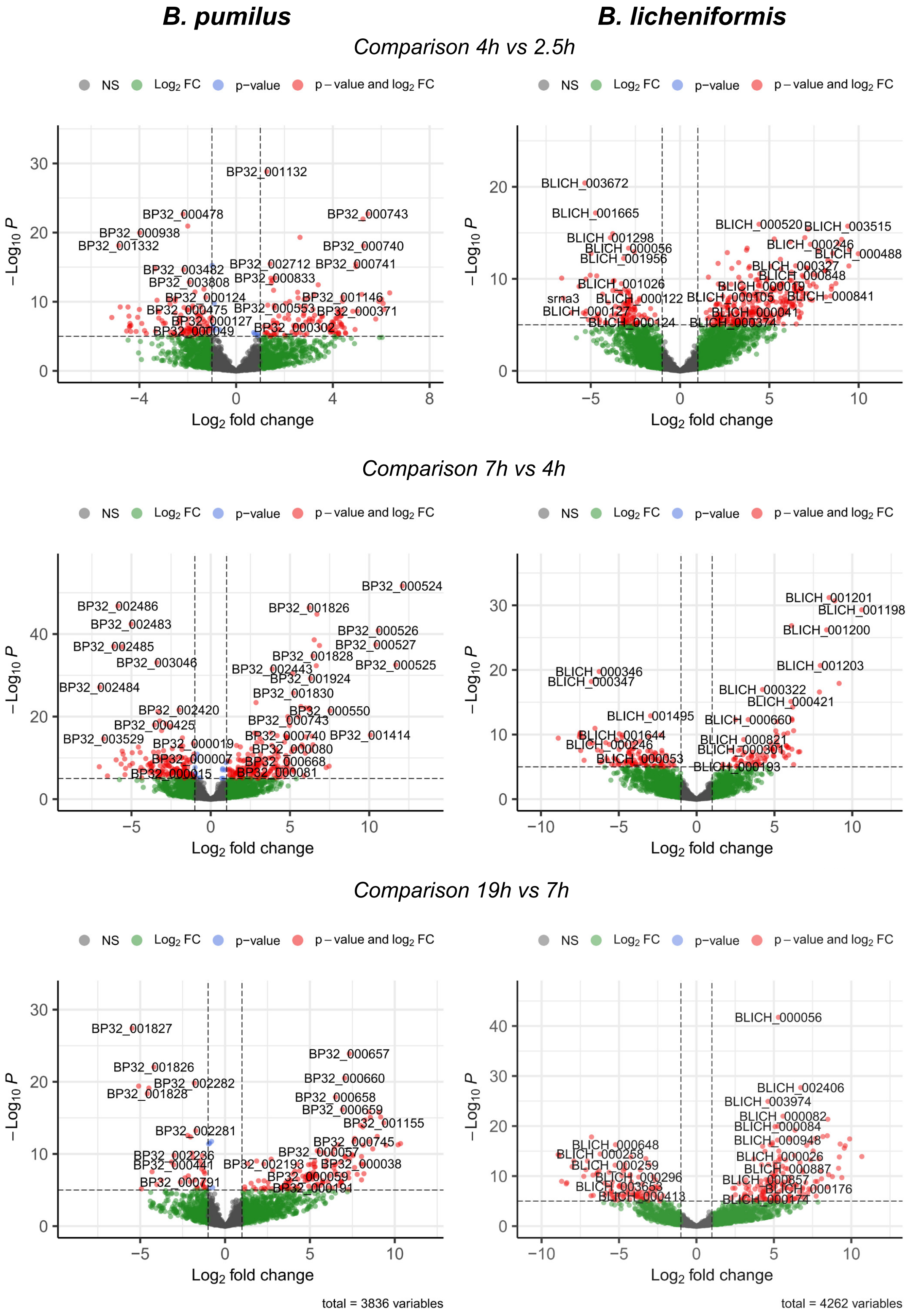


FIGURE 3.16: Volcano plots representing the DEG analysis for *B. pumilus* MS32 and *B. licheniformis* MW3 Δ *yqfD* at different fermentation timepoints comparisons. The cutoffs for p-value and log2FC are 10e-6 and 2, respectively. Genes passing both thresholds are depicted in red.

These comparisons highlight genes with relevant roles associated to the different stages of the fermentation. For example, in *B. pumilus* MS32 at the transition phase (4 hours) several genes are significantly regulated in comparison to the first sampling point (2.5 hours). The top most significantly regulated genes include those coding for: oligoendopeptidase F, WD40 repeat domain-containing protein, and a stress protein.

3.4.6 Specialized products

Insights from the comparative genomic analysis by Proteinortho [188], Antismash [30] and COGClassifier [295] were complemented by the transcriptional information generated by DP_GP_cluster and Annogesic [363]. For example Table 3.17 presents the combined results regarding the third BGC region identified by Antismash [30]. According to Proteinortho [188], these biosynthetic genes are absent in *B. subtilis* and *B. licheniformis*. The transcriptional profiles of the core and accessory biosynthetic genes are depicted together with the cluster and putative target sRNAs.

TABLE 3.17: Characterization of the biosynthetic gene cluster for the synthesis of a NRP-Polyketide in *B. pumilus* MS32. TPM corresponds to triplicate mean.

Locus Tag	Product	Cluster	TPM 2.5	TPM 4	TPM 7	TPM 19	Interacting srna
BP32_000675	amino acid adenylation domain-containing protein	16	98.2	958.5	408.4	143.31	NA
BP32_000676	beta-lactamase family protein	16	127.7	1274.0	458.4	121.68	srna2, srna21, srna32, srna33
BP32_000677	C39 family peptidase	16	104.9	953.1	323.9	80.52	srna29
BP32_000678	alpha/beta fold hydrolase	16	109.7	1064.6	332.0	91.88	srna16, srna 25, srna35
BP32_000679	3-hydroxyacyl-CoA dehydrogenase family protein	7	146.1	1457.5	488.0	116.91	srna4, srna20, srna29, srna32, srna34
BP32_000680	HAD-IIIC family phosphatase	16	141.2	1400.8	380.7	113.37	srna24, srna25, srna33
BP32_000681	acyl carrier protein	16	113.5	1269.4	400.1	96.49	NA
BP32_000682	acyl-CoA dehydrogenase family protein	16	111.3	1376.6	431.2	106.77	srna13
BP32_000683	amino acid adenylation domain-containing protein	7	64.9	1504.8	465.2	100.07	NA
BP32_000684	amino acid adenylation domain-containing protein	7	49.4	1254.4	339.8	51.62	srna3, srna9, srna29, srna32, srna33
BP32_000685	SDR family NAD(P)-dependent oxidoreductase	16	51.1	1247.2	319.6	62.83	srna1, srna3, srna8, srna15, srna17, srna24, srna30, srna31
BP32_000686	SDR family NAD(P)-dependent oxidoreductase	16	50.1	1226.9	371.1	75.36	srna4
BP32_000687	HAD-IIIC family phosphatase	9	50.1	1169.4	407.1	97.31	srna4
BP32_000688	phosphotransferase enzyme family protein	9	117.4	1456.3	897.6	635.43	NA

Most of the BGS genes for the region 3 identified by Antismash [30] follow the transcriptional activity depicted by clusters 16 and 7, moreover these transcripts are remarkably abundant at the transition point (4 hours) and are potentially regulated by several candidate sRNAs. Some of these sRNAs were identified as similar to those identified in *B. altitudinis* SCU11, specifically srna32, srna13, srna3 and srna17 which correspond to Bpsr1, Bpsr92, Bpsr178, and Bpsr70, respectively.

Similarly, a characterization for the genes encoding proteins related to Iron transport are presented in Table 3.18. Those proteins were identified in *B. pumilus* MS32 but not in *B. subtilis* 168 or *B. licheniformis* according to the Proteinortho [188] analysis.

TABLE 3.18: Characterization of Iron transporters unique to *B. pumilus* MS32. TPM corresponds to triplicate mean.

Locus Tag	Cluster	Product	TPM 2.5	TPM 4	TPM 7	TPM 19	Interacting sRNA
BP32_003238	2	iron-hydroxamate ABC transporter substrate-binding protein	240.58	277.31	96.54	120.23	srna9, srna12
BP32_003239	2	iron ABC transporter permease	167.31	236.29	116.55	120.55	NA
BP32_003240	16	iron ABC transporter permease	117.19	161.22	117.57	120.31	NA
BP32_003244	9	ABC transporter substrate-binding protein	37.36	115.20	69.78	49.91	srna38
BP32_003660	7	heme oxygenase	190.16	586.83	357.71	48.04	srna7, srna11, srna36
BP32_003662	7	ABC transporter substrate-binding protein	208.27	750.70	443.83	68.63	srna0
BP32_003664	7	ABC transporter ATP-binding protein	196.40	634.23	388.76	67.84	srna4
BP32_003665	7	iron ABC transporter permease	195.29	799.20	559.15	65.92	srna3
BP32_003666	7	heme ABC transporter substrate-binding protein IsdE	222.35	715.47	510.60	61.89	NA
BP32_003667	7	NEAT domain-containing protein	434.42	1409.99	1311.95	157.86	NA
BP32_003668	7	heme uptake protein IsdC	401.29	1289.50	1292.56	165.32	NA
BP32_003669	7	NEAT domain-containing protein	481.50	1443.38	1540.26	199.17	NA

Table 3.18 shows the genes coding for two NEAT domain-containing, and heme uptake proteins are highly transcribed at the 4 and 7 hours sampling points, these transcripts do not appear to be targeted by any of the candidate sRNAs. Most of the genes presented in the table are members of the DP_GP_cluster 7 and reached maximal transcriptional activity at the transition point. Potential interacting sRNAs include srna0 and srna3, which seem related Bpsr193 and Bpsr178 of *B. altitudinis* SCU11.

3.4.7 Proteases and protein secretion

Tables 3.19 and 3.20 provide a detailed characterization of the transcriptional activities of the main proteases of *B. pumilus* and *B. licheniformis*. Cluster membership, putative interacting sRNAs and triplicate mean TPM values are presented.

TABLE 3.19: Characterization of *B. pumilus* proteases. Transcriptional activities per time point, corresponding DP_GP cluster, cellular location, and predicted interacting sRNAs for some relevant proteases of *B. pumilus* MS32. TPM=Transcripts per million, reported as triplicate mean. E=Extracellular, EC=Extracellular wall associated, C=Cytosolic, M=Membrane.

Protease	LocusTag	Location	Cluster	TPM 2.5	TPM 4	TPM 7	TPM 19	Interacting sRNA
AprE	BP32_001020	E	10	2.5	111.1	370.3	1950.2	srna12
Bpr	BP32_001477	E	10	10.7	14.3	321.4	1144.9	srna28
Epr	BP32_000293	E	4	157.3	159.1	111.5	104.8	srna6
Mpr	BP32_001755	E	10	31.0	146.6	234.0	649.4	srna12
Vpr	BP32_003595	E	17	114.2	263.1	231.5	314.2	srna2, srna28
Wrpa	BP32_001270	EC	10	18.8	163.6	743.6	4378.1	NA
AprX	BP32_001682	C	14	0.4	1.0	1.1	221.1	srna18, srna20
Isp	BP32_002298	C	4	160.7	160.7	75.5	76.4	srna8, srna20, srna34
FtsH	BP32_000085	C	2	1504.1	1949.4	864.1	873.0	NA
MlpA	BP32_001640	C	3	142.6	122.6	45.8	88.7	NA
SppA	BP32_002723	M	16	279.5	601.2	233.3	171.3	srna2, srna15, srna23, srna41
CtpA	BP32_001959	M	13	243.9	177.0	127.3	111.7	srna37
CtpB	BP32_003304	M	14	5.1	4.9	4.9	48.6	srna24
HtpX	BP32_001295	M	18	353.5	398.3	360.2	164.7	srna19, srna20

TABLE 3.20: Characterization of *B. licheniformis* proteases. Transcriptional activities per time point, corresponding DP_GP cluster, cellular location, and predicted interacting sRNAs for some relevant proteases of *B. licheniformis* MW3. TPM=Transcripts per million, reported as triplicate mean. E=Extracellular, EC=Extracellular wall associated, C=Cytosolic, M=Membrane.

Protease	LocusTag	Location	Cluster	TPM 2.5	TPM 4	TPM 7	TPM 19	Interacting sRNA
AprE	BLICH_001097	E	11	1.7	0.2	2.6	3175.2	NA
Bpr	BLICH_001738	E	16	5.1	16.4	622.2	4198.1	srna0, srna1, srna3, srna26, srna31, srna62
Epr	BLICH_001111	E	17	600.4	1031.6	230.3	32.2	NA
Mpr	BLICH_000338	E	10	97.8	23.5	261.0	3447.5	srna37
Vpr	BLICH_003976	E	10	96.4	19.8	453.7	1138.1	NA
Wrpa	BLICH_002831	EC	10	51.7	19.9	129.1	503.9	srna23, srna44, srna74
AprX	BLICH_002170	C	11	2.3	1.9	1.9	4.1	srna70
Isp	BLICH_002583	C	2	197.9	67.7	140.7	113.8	srna27
FtsH	BLICH_000087	C	5	1449.9	772.1	1085.6	685.1	NA
MlpA	BLICH_001885	C	1	365.0	87.0	153.9	137.3	NA
SppA	BLICH_003074	M	1	237.2	98.2	84.2	86.9	srna55, srna58
CtpA	BLICH_002266	M	22	262.7	133.6	302.7	225.8	Na
CtpB	BLICH_003724	M	20	4.3	39.6	22.1	42.9	srna50
HtpX	BLICH_001495	M	15	297.7	1171.9	224.5	152.3	srna62

Regarding the extracellular proteases, as expected, most of them reached peak transcriptional activity at the late stages of the fermentation (Tables 3.19 and 3.20). The exception was *epr*, which for both *Bacillus* was most transcribed during the transition point. In *B. subtilis*, Epr is part of SinR, ScoC, Spo0A, DegU and SigD regulons and plays a role in cell-to-cell communication and regulation of swarming mediated by DegU [132]. Therefore, it makes sense to find it highly active at 4 hours in the fermentation, when coordination of the cell culture and determination of cell fates occurs.

There are eight extracellular proteases encoded by *B. subtilis*. Homologous genes encoding those proteins were found in *B. pumilus* MS32 and *B. licheniformis* MW3, except for NprB and NprE. In *B. subtilis*, AprE together with NprE are the major proteases and are accounted for around 95% of the extracellular proteolytic activity [132]. In contrast, according to the transcriptional activities during the fermentation runs, AprE and Wpra appeared as the major extracellular proteases for *B. pumilus* MS32 (Table 3.19), while Bpr and Mpr are the most prominently transcribed proteases of *B. licheniformis*, followed closely by AprE (Table 3.20).

Interestingly, transcripts for the wall associated WprA were highly abundant for *B. pumilus* MS32, with 2.2 times more TPM than AprE (Table 3.19). No sRNA candidate was predicted to interact with *wrpA* mRNA in *B. pumilus*, while three putative sRNAs might target it in *B. licheniformis* (Tables 3.19 and 3.20). It is worth of notice that despite identification of the known aprAs sRNA in *B. licheniformis* [139], the known interaction with AprE was not identified. True biological targets with low scoring by prediction algorithms is a known issue of computational methods [245], it is possible that the interaction between *aprE* mRNA and the corresponding antisense sRNA ranked below the scoring threshold used in Annogesic.

Tables 3.19 and 3.20 show divergences in the transcriptional profiles of the main proteases of *B. pumilus* and *B. licheniformis*, the putative interacting sRNAs for ech of them also differ. The biological roles of these proteases and differences will be further discussed in the following chapter.

The transcriptional profiles of main *Bacillus* secretory elements and their potential regulators were compared in order to gain understanding on when and how these features are active and interacting during the course of a fermentation. Tables 3.21 and 3.22 present a characterization of main secretory components and the putative interacting sRNAs of *B. pumilus* MS32 and *B. licheniformis* MW3.

TABLE 3.21: Characterization of secretory machinery and accessory components of *B. pumilus* MS32. Transcriptional activities per time point, corresponding DP_GP cluster and predicted interacting sRNAs. TPM=Transcripts per million, reported as triplicate mean.

Gene	LocusTag	Cluster	TPM 2.5	TPM 4	TPM 7	TPM 19	Interacting sRNA
secA	BP32_003315	9	439.1	733.7	616.1	548.1	srna5, srna21
secE	BP32_000131	10	1264.8	1404.1	1530.9	1812.2	NA
secDF	BP32_002549	2	367.0	580.7	276.6	280.2	srna30, srna31
secG	BP32_003186	5	355.8	494.7	363.3	450.1	srna28, srna37
secY	BP32_000167	1	8534.7	3273.8	562.4	426.8	NA
ftsY	BP32_001548	4	233.8	187.4	118.5	94.1	srna27, srna33
dnaK	BP32_002420	4	1788.2	2015.5	356.8	334.7	NA
psrA	BP32_000989	16	692.0	1075.7	654.1	685.3	NA
groEL	BP32_000606	4	3281.4	5233.4	1152.7	915.2	srna34
grpES	BP32_000605	4	3721.7	5280.6	872.1	739.1	NA
floA	BP32_002411	8	384.6	260.2	169.0	103.9	srna27
floT	BP32_002894	8	175.7	102.4	66.5	23.4	srna39
tatC	BP32_000601	6	380.6	967.2	3014.5	1767.5	srna14, srna42
tatAE	BP32_000600	17	364.0	770.4	2140.8	2118.2	srna1, srna2, srna14, srna15, srna30, srna31, srna35, srna42

TABLE 3.22: Characterization of secretory machinery and accessory components of *B. licheniformis* MW3. Transcriptional activities per time point, corresponding DP_GP cluster and predicted interacting sRNAs. TPM=Transcripts per million, reported as triplicate mean.

Gene	LocusTag	Cluster	TPM 2.5	TPM 4	TPM 7	TPM 19	Interacting sRNA
secA	BLICH_003732	15	560.8	2671.1	826.5	468.0	srna18, srna43
secE	BLICH_000121	10	398.8	164.3	423.8	823.5	NA
secDF	BLICH_002875	2	373.0	119.4	235.7	207.9	srna1
secG	BLICH_003609	13	699.6	208.9	526.3	917.0	srna12
secY	BLICH_000156	6	3976.6	1718.0	542.1	398.0	NA
ftsY	BLICH_001805	17	176.1	135.3	172.4	99.7	srna1
dnaK	BLICH_002720	18	1025.0	1485.2	616.2	401.4	NA
psrA	BLICH_001061	22	631.1	420.4	695.1	641.1	srna48
groEL	BLICH_000621	5	3097.1	1747.4	1913.1	1385.8	srna29, srna49
groES	BLICH_000620	5	2397.9	972.8	1289.2	777.2	srna67, srna72
floA	BLICH_002710	15	410.5	3178.6	539.3	216.5	srna68
floT	-	-	-	-	-	-	-
tatC	BLICH_000617	8	432.9	2528.0	1532.6	1542.0	srna25, srna32, srna35, srna51, srna54, srna61
tatAE	BLICH_000616	8	536.0	4507.5	2358.9	2919.6	srna32, srna35, srna51, srna61

According to the predictions by Annogesic [363], several candidate sRNAs target the different components of the secretory machinery in both *B. pumilus* and *B. licheniformis*. Regarding the Sec apparatus, no sRNA seem to target mRNAs of *secE* and *secY* in the strains studied. This could point to regulation occurring a at different layer or by so far unidentified interaction partners missed by the detection parameters of this analysis.

Tables 3.21 and 3.22 depict similarities for *B. pumilus* and *B. licheniformis* regarding secretory components. For example in both species maximal transcriptional activity for *secA* and *secE* occurred at 4 and 19 hours, respectively, while *secY* and *ftsY* transcripts were more abundant at 2.5 hours. In contrast some genes showed differences in their transcription profiles, for *B. pumilus secDF* and *secG* showed maximum transcription at the transition point, whereas in *B. licheniformis secDF* and *secG* reached peak activities at 2.5 and 19 hours, respectively. Additionally, *B. pumilus tatC* and *tatAE* transcripts were more abundant at 7 hours, while in *B. licheniformis* these genes were most transcribed at the 4 hours. This points to some inter-species differences regarding transcription and regulation of secretory components.

Interestingly, *tatC* and *tatAE* mRNAs shared potential interacting sRNAs in both *Bacillus* species (Tables 3.21 and 3.22). In *B. pumilus* the candidates srna14 and srna42 were predicted to interact with these two transcripts. For *B. licheniformis*, the putative sRNAs targeting both *tatC* and *tatAE* are even more (srna32, srna35, srna51 and srna1).

3.4.8 Transcriptional activities of key regulators

Several master regulators coordinating global bacterial responses and key metabolic processes have been identified in *B. subtilis* and collected in resources such as SubtiWiki [250]. These regulatory elements were also identified in *B. pumilus* MS32 and *B. licheniformis* MW3 by PGAP [312] annotations and Proteinortho [188] analysis. Tables 3.23 and 3.24 present some crutial regulators, their assigned transcriptional profile by DP_GP cluster [220], transcriptional activities and putative interacting sRNAs.

TABLE 3.23: Characterization of key regulators in *B. pumilus* MS32. Transcriptional activities per time point, corresponding DP_GP cluster and predicted interacting sRNAs. TPM=Transcripts per million, reported as triplicate mean.

Gene	LocusTag	Cluster	TPM 2.5	TPM 4	TPM 7	TPM 19	Interacting sRNA
abbA	BP32_001362	10	283.7	317.5	610.6	1056.2	NA
abrB	BP32_000051	4	237.9	310.1	44.2	25.3	NA
ccpA	BP32_002760	6	127.7	224.5	563.8	253.7	srna13
codY	BP32_001569	4	456.5	438.9	189.0	208.1	NA
csrA	BP32_003323	9	101.3	354.9	445.1	306.8	NA
scoC	BP32_000994	7	101.3	164.1	318.0	91.8	NA
spo0A	BP32_002292	10	241.8	480.8	1104.0	1066.9	srna20, srna23, srna29, srna30, srna31, srna40
swrA	BP32_003303	17	108.3	826.3	1617.5	1130.3	srna30, srna31
swrB	BP32_001601	18	205.2	242.6	187.7	103.6	srna12, srna39, srna42
degU	BP32_003335	9	1608.6	2588.2	2535.7	2211.4	srna18
degS	BP32_003336	5	238.0	330.1	193.4	406.3	srna1, srna17
degQ	BP32_003016	10	149.6	333.9	410.8	822.5	NA
sinR	BP32_002331	17	215.2	326.4	402.5	450.0	srna36
sinI	BP32_002330	17	26.5	115.7	140.2	146.0	srna10
slrR	BP32_003236	6	8.1	9.1	11.1	8.6	NA
slrA	BP32_003611	14	60.6	62.9	73.2	140.5	srna19
comA	BP32_003012	11	179.6	220.7	159.3	300.9	srna0, srna3, srna17

TABLE 3.24: Characterization key regulators in *B. licheniformis* MW3. Transcriptional activities per time point, corresponding DP_GP cluster and predicted interacting sRNAs. TPM=Transcripts per million, reported as triplicate mean.

Gene	LocusTag	Cluster	TPM 2.5	TPM 4	TPM 7	TPM 19	Interacting sRNA
abbA	BLICH_001618	10	87.4	74.2	433.3	705.2	srna13, srna25
abrB	BLICH_000053	18	793.0	1211.7	145.3	114.2	srna34, srna43
ccpA	BLICH_003103	13	178.6	79.0	140.4	162.4	srna22, srna46
codY	BLICH_001826	14	448.8	166.0	145.9	211.3	NA
csrA	BLICH_003740	17	362.6	379.6	232.3	36.6	NA
scoC	BLICH_001067	4	785.6	525.9	982.8	500.3	NA
spo0A	BLICH_002577	16	263.8	506.3	632.5	1232.9	srna24, srna33, srna68
swrA	BLICH_003723	17	271.7	195.7	81.1	22.5	srna28, srna71
swrB	BLICH_001858	6	334.1	86.4	52.0	28.1	NA
degU	BLICH_003752	20	333.5	815.8	970.6	1086.1	srna8, srna39, srna40, srna45, srna50, srna65, srna68
degS	BLICH_003753	7	137.5	176.9	192.8	140.6	srna19, srna32, srna65, srna70
degQ	BLICH_003329	11	32.6	42.5	27.0	73.1	srna31
sinR	BLICH_002619	16	96.0	115.4	159.7	311.9	srna51, srna61
sinI	BLICH_002618	3	4.3	12.6	10.5	44.6	srna40, srna47, srna51, srna61
slrR	BLICH_003658	17	29.9	23.9	25.0	0.0	srna62
slrA	BLICH_003998	25	47.1	96.2	50.5	14.7	srna43, srna64
comA	BLICH_003323	15	72.4	83.4	71.3	72.4	srna12, srna29

Tables 3.24 and 3.23 allow a comparison between components of essential regulatory networks of *B. pumilus* MS32 and *B. licheniformis* MW3. For both species, no candidate sRNA seem to target *codY*, *csrA*, *scoC* mRNAs while *spo0A* transcripts appear as highly targeted for regulation by sRNAs. A remarkable difference regarding the two component system response regulator DegS-DegU was observed. The transcriptional activity of *degU* is notably higher than any other regulator within the set in *B. pumilus* MS32, and seems targeted only by srna18. In contrast, *B. licheniformis degU* appears as highly regulated by seven candidate sRNAs and its transcription levels were in a comparable range with the other regulators analyzed. The functions and relevance of these regulators will be further explored in the following chapter.

Chapter 4

Discussion

A multi layer comparative systems analysis study was performed to assess the potential of *B. pumilus*, a close relative to *B. licheniformis* and *B. subtilis*, as an emerging industrial production host. The first stage of the study entailed a comprehensive genomic characterization of the novel isolate *B. pumilus* MS32 in regards to other described *B. pumilus* strains, and a detailed inter-species genomic comparison with *B. licheniformis* and *B. subtilis*, two species known to offer excellent workhorses for biotechnological applications. Once the genomic potential of *B. pumilus* was investigated, the next step consisted in creating a collection of germination deficient mutants to be run in small scale fermentations. Samples from relevant time points of the process were collected, which opened the door to characterization at different growth stages. A RNA isolation protocol was optimized for this type of samples and high-quality RNA was purified. The RNA was further processed to create RNA-seq libraries from which transcriptomic studies were conducted. Transcriptional activity data was interrogated from multiple perspectives, including differential expression analysis, transcription profiling across time points and prediction of potential regulatory small RNAs and their putative interacting partners. The analysis was focused on features impacting cell responses and key process of interest such as protein secretion and RNA-mediated regulation. Such investigations facilitate and complement the understanding of these *Bacillus* species biology, and the discovery and characterization of targets for optimization of microbial cell factories.

4.1 Germination deficient mutants, a case of genetic accessibility

Spore-forming bacteria in an industrial set up involves higher sterilization costs and potential risk of contamination, as some spores from *Bacillus* species are known to be particularly resilient [169, 325]. Furthermore, the production of spores is an energy demanding process for the cell, and when triggered, it consumes resources that otherwise could be used for production and secretion. As spores are metabolically dormant, their presence decreases productivity. Therefore, sporulation is a feature commonly engineered out of *Bacillus* strains used in industry [236, 373]. In order to perform small scale fermentations at the research and development facilities of the industrial partner, one of the first goals of this project was to generate a collection of *B. pumilus*, *B. licheniformis* and *B. subtilis* mutants unable to produce viable spores.

As shown in the previous chapter, while transformation of *B. subtilis* 168 to produce a germination negative mutant was simple and straightforward to achieve, the same could not be accomplished for the type strain *B. pumilus* DMS27. Protocols designed and optimized for *Bacillus licheniformis* and *B. subtilis* were not suitable for DSM27. This relates to species and even strain specific properties of each bacteria. Generally, to transform a wild-type strain represents a bigger challenge than modification of a domesticated laboratory bacterium, and entails many rounds of optimization and labor-intense experiments. However, it is necessary in order to gain access to their promising characteristics and enable genetic engineering for production optimization.

4.1.1 Barriers to genetic accessibility

There are previous reports already indicating that certain *B. pumilus* strains are recalcitrant in matters of genetic manipulations. Research shows that a method developed for one strain might not be suitable for another one [291, 344, 70, 50]. The experiments here described for the type strain *B. pumilus* DSM27 add to these observations. A study on the strain Jo2 reported that protoplast transformation was hardly achievable and not possible when PCR fragments were used [344]. Recently, a stable resistance to accept foreign DNA was also observed for the strain 3-19, even when an electroporation protocol successfully optimized for the strain DX01 was applied; it was only after testing different parameters that transformation could be achieved [70]. Similarly, the method described for the DX01 strain failed to generate knock-out mutants for the strain SAFR-032 [50].

The thick cell wall of Gram positive bacteria constitutes one physical barrier for transformation. The composition and arrangement of the cell wall varies among species, strains, and even physiological conditions, which complicates to make generalizations on the best approach to overcome the barrier. For example, it has been reported that glycine and DL-threonine used as cell-wall-weakening agents can enhance the transformation efficiency in electro-competent cells [361]. By replacing the L and D-alanine bridges, the incorporation of these amino acids reduce the cross-linking of the peptidoglycan layer, which in turn makes the cell wall more loose and DNA entrance easier [367, 361]. Nevertheless, while this approach increased electroporation efficiency in *B. subtilis* and *B. amyloliquefaciens*, for *B. cereus* it had the opposite effect, differences in cell wall compositions were offered as a possible explanation.

Each transformation method provides a different approach to deliver DNA into the cell. While protoplast transformation relies on enzymatic digestion of the cell wall, electroporation produces transient pores due to cell exposure to high-voltage pulses. On the other hand tribos transformation applies "nano-needles" to pierce the bacteria, and conjugation makes use of a sophisticated mating apparatus which results in a pore through which DNA is transported [15]. These different techniques, which turned successful in many other bacteria, failed to produced the desired *B. pumilus* DSM27 germination negative mutant, therefore, other barriers must be in play.

Assuming that all tested methods successfully delivered DNA into *B. pumilus* DSM27, the next hurdle for efficient incorporation of DNA into the host genome must be within the cell. RM systems are widespread among Bacteria and Archea organisms, and represent a major barrier to genetic manipulation [339], particularly when the bacterium of interest carries multiple RM-systems [368]. According to the analysis against the REBASE database [274], that is the case for several *B. pumilus* strains (Table 3.3).

The failed transformations suggest that *B. pumilus* DSM27 carry very active RM systems, by which the deletion cassette designed for *yqfD* is degraded before recombination events can occur. If active degradation of the DNA material by DSM27 endonucleases is the main obstacle for genetic accessibility, there are a couple of strategies to circumvent the problem, either by changing the methylation pattern of the exogenous DNA or by reducing the restriction activity of targeted bacteria [15].

In some cases a heat-shock step allows to temporarily inactivate the cell's RM systems, such was the objective of the incubation at 55 °C step in the tribos protocol. Inactivation of host RM systems by heat treatment was reported to increase transformation efficiency in recalcitrant *B. amyloliquefaciens* strains [367], however another study showed that in *B. subtilis* ZK the heat-shock decreased the transformation efficiency [371]. Since heat sensitivity of the restriction enzymes can vary between organisms, the step must be tailored in a case-basis.

Another approach to evade the activity of host endonucleases is by altering the methylated state of the incoming DNA. A simple method could be incubation of the DNA with crude cell extracts of the host, so the methyltransferases could modify it prior to transformation [123], with the drawback of endonucleases also present that might act before methylation occurs. Alternatively, DNA can be transformed into an intermediate bacteria with a different methylation profile, producing a signature that evades the host restriction enzymes [185]. Another option is to engineer the intermediate bacteria to mimic the methylation of the targeted organism [358]. Finally, commercially available methyltransferases can be used to treat the transforming DNA [185, 309]. The latter was reported as an effective strategy for the transformation of the a-amylase producing *B. amyloliquefaciens* Z3 [309].

Tackling out the RM systems does not guarantee optimal transformability for the bacteria under study. Acquisition, processing and incorporation of exogenous DNA is a multi-layer process under sophisticated regulation, and therefore, other barriers could interfere with transformation efficiency. For example, in the development of *B. licheniformis* MW3, even after the deletion of genes encoding type 1 restriction enzymes, the strain still presented low transformation efficiency [140], as the study describes, elements and regulators of competence and quorum sensing processes might render low transformation efficiencies when they are poorly interlinked. Even more, activity of mobile genetic elements can also relate to the ability of bacteria to uptake DNA [140].

Systematic optimization is labor-intensive given the wide array of parameters to consider, some of which interact between each other adding another level of complexity to the experiments. To try those alternatives and more variations of the methods described in literature would have been time consuming. Moreover, comparative genomic analysis of *B. pumilus* revealed an unusually high number of RM systems present in the strain DSM27 (Table 3.3), which offers a possible explanation for its resistance to transformation. The analysis also pointed to *B. pumilus* MS32 (a strain already genetically accessible) as a close strain to DSM27 and suitable candidate to substitute it for downstream experiments.

4.2 Comparative genomic insights

One crucial step in comparative genomic experiments is annotation, for this project PGAP [312] was selected for this task. There are some alternatives for bacterial whole genome annotation, such as Prokka [290] and RAST [16], which offer a good balance between quality and speed of annotation, achieved by using smaller curated databases. A potential drawback of those approaches is that proteins from divergent or novel genomes might be left without annotation [279]. In contrast, tools like PGAP [312] make use of extensive and interconnected databases supported by International Nucleotide Sequence Database Collaboration (INSDC), which is more time and memory intensive, but achieves a superior annotation quality.

Average Nucleotide Identity. According to general convention, two genomes of the same species tend to share more than 95% identity of ANI [118]. Notably, the ANI values for the *B. pumilus* strains SH B11, MTCC B6033, and TUAT1 fall below 89%, which is under the boundary delimiting a bacterial species Figure (3.1). This is not surprising, as species delimitation within closely related *Bacillus* often requires careful evaluation beyond 16S rRNA sequences; more comprehensive approaches include MLST (MultiLocus Sequence Typing) [157, 208] and full genomic comparisons. A study characterizing *B. pumilus* isolates from terrestrial and marine environments found those strains to be closer to the *B. altitudinis* a species closely related to *B. pumilus*. [108].

A recent study introduced *B. pumilus* HM-7, which is associated to bacterial soft rot in melon [337]. The authors delineated evolutionary relationships among 21 *Bacillus pumilus* strains and found that HM-7 belongs to a clade which included the strains MTCC-B6033, TUAT1, C4, SH-B11, and GR8, some of which were found distant to MS32 and the type strain DSM27 by this study. *B. pumilus* GR8 has known pathogenic effects to ginger [253]. These observations are important to assess the bio safety of bacterial organisms used for industrial applications, particularly to recognize potential risks and keep track of origin, acquisition, and evolution of pathogenic traits within a group of interest.

Insertion Sequence Elements. Despite the contributing role of IS to rapid evolution, genome plasticity, and trait acquisition; activation of these elements can result detrimental for a productive process [235, 252]. Within a fermentation, several aspects can induce IS activation (temperature, oxidative stress, host factors, antibiotics). Once active, IS elements can mediate gene (in)activation and mobilization, alter the expression of adjacent genes, and produce topological changes in DNA, these changes might impact productivity features of an industrial strain [328, 53]. Genetic instability can cause a bacterial population to loose desirable morphological and biosynthetic characteristics, this is known as strain degeneration, and is an issue of concern for industrial production [252]. For example, in *B. subtilis* natto, a strain synthesizing poly-γ-glutamic acid (γPGA) during natto (fermented soybean) production, it was reported that translocation of the IS4Bsu1 element into the *swrA* gene impaired the γPGA synthesis capability of the strain [172, 235]. Moreover, engineering of IS-element free bacteria has been proposed as an alternative to generate stable and efficient host strains for both laboratory and industrial applications [252, 53]. Detection of IS within *B. pumilus* genomes might aid in identification of potential optimization targets.

Prophages. Interactions of phages with their host bacteria are subjected to co-evolution forces and can lead to emergence of advantageous features, for example: immunity-related proteins could protect the host from further foreign phage infections [7]. Phages can also encode features related to competitive advantage and resistance mechanisms that enhance host survival [283]. However, the phage-host interactions can also have detrimental effects, for example in food industry the phage infection of a starter bacterial culture, such as *B. subtilis*, represents a serious contamination concern [273]. When a prophage adopts a lysogenic lifestyle, the genes related to cell lysis are repressed and the prophage replicates within the host genome, but upon prophage induction they become active and induce lysis of the host cell [46]. If a prophage changes to a lytic lifestyle during a fermentation, the total cell biomass is affected, this reduction in cell density might lead to a negative effect on product yield, not only by the decrease in productive cells but also due the proteases released by the cell lysis, which could degrade the secreted products [338]. A study on *B. subtilis* showed that by deletion of genes associated with cell lysis (one of them the *xpf* from the PBSX prophage) increased in biomass and in production of recombinant enzymes was obtained. Prophages represent another optimization point for development of microbial cell factories.

CRISPR-Cas systems. Interestingly, a recent study on 1871 genomes from the *B. cereus* group, observed that inactivation of CRISPR-Cas systems correlates with higher acquisition of mobile genetic elements (MGEs) [372]. When the authors compared genomes with and without active CRISPR-Cas systems, they observed that bacteria with functional CRISPR-Cas were limited to more specific niches, had less MGEs and less unique genes. For these bacteria, CRISPR-Cas systems seem to act as a barrier to horizontal gene transfer, limiting the acquisition of genetic traits that could be beneficial for adaptation to diverse environments. Alternative defense mechanisms, like Restriction-Modification systems, might takeover the role to protect the cells against phage infections when CRISPR-Cas systems are inactive [372]. Considering the metabolic versatility and wide range distribution of *B. pumilus*, the abundance of prophages and IS elements in the analyzed genomes, together with the seldom CRISPR-Cas predictions (Table 3.2), it appears that the observations of the study could also apply to *B. pumilus*, of course the analysis should be expanded to more genomes and to identification of more MGEs before giving a conclusion.

Additionally, it is relevant to consider that the role of CRISPR-Cas systems in limiting gene flow remains under debate [117, 372, 276]. For example, it has been proposed as well that an active CRISPR-Cas could facilitate gene flow during transduction events. In such cases, a bacterial cell would be protected by the CRISPR-Cas system targeting the phage, but at the same time it would receive additional exogenous DNA also packed within the viral particle[276, 340].

Restriction Modification systems. As discussed previously, bacterial RM systems constitute a significant barrier against genetic manipulation. Their protective role against invading phages in absence of active CRISPR-Cas was also highlighted. Nevertheless, there is increasing evidence pointing to additional roles of RM systems in bacteria which could also be in play for the studied *Bacillus*.

Functions beyond defense have been proposed as an explanation for the high abundance, wide distribution, large specificity range, and independent evolution of endonucleases regarding methyltransferases [329, 154], as they represent characteristics difficult to explain if RM systems are limited only to cell defense. For instance, RM systems have been described as selfish genetic elements [176], they have also been associated to stabilization of genomic islands, reallocation of deoxyribonucleotides for viral DNA production, and to generation of additional substrates for the recombination machinery [329]. Furthermore, the presence of multiple RM systems with different recognition patterns within members of a species, also points to a mechanism for genetic isolation and maintenance of species identity. According to this, bacteria from the same species carry distinctive signatures (almost like a barcode) produced by RM systems, separating them into strains that would not interchange genetic material, over evolutionary time, new species could emerge from such variants [154].

Overall, it is necessary to consider these highly dynamic elements within an actively evolving network, where a mobile element might be interacting with a bacterial host as well as with other MGEs (both resident and incoming). The attack/defense mechanisms, both from the bacterial or the selfish element side are co-evolving as complex multilayered systems in a ever going battle for survival. As recently pointed, novel strategies, systems and mechanisms are just being discovered, and there is vast landscape open for exploration [276].

Secondary metabolites. The production and characterization of the secondary metabolites identified in *B. pumilus* MS32 (Table 3.4) remain to be confirmed by experimental approaches. Among the known clusters, MS32 could potentially synthesize compounds similar to lichenisyn, bacilysin, bacillibactin, fengycin and plantazolicin, which have been already characterized in related *Bacillus* species.

For example, produced by *B. licheniformis*, the highly stable surfactant lichenysin is an amphiphilic lipopeptide with applications such as biocontrol in agricultural industry, oil emulsifier, and foaming agent in cosmetics. Moreover it shows good ion chelating properties and antibiotic activity. Despite its applications, production of lichenysin is of concern in food processing [205, 262]. In a recent study it was found that lichenysin synthesis capacity by 11 strains of *B. licheniformis* up to 5log10 viable cells/ml in liquid food was unlikely to constitute a risk, nevertheless, upon favorable conditions cell density might rise and production of lichenysin could reach concentrations that trigger foodborne intoxication, therefore, it was recommended to monitor and prevent the production of this compound along the food chain [359].

Bacilysin is a 270 Da dipeptide antibiotic non-ribosomally produced by many *Bacillus* species. Despite its simple structure, bacilysin presents antagonistic activity against a broad spectrum of fungi, algae and bacteria, and therefore, is of biotechnological interest due to its antimicrobial applications in bio-preservation [237, 150]. For *B. pumilus*, the antimicrobial peptide was early named as tetaine, but it was later found to be chemically and physically identical to bacilysin [244, 160]. A recent study characterizing the bacilysin gene cluster within the *B. subtilis* group, found that it was incomplete for *B. pumilus* isolates, missing a gene coding for a responsible for export and intrinsic resistance [237]. The authors suggested that a yet unknown permease or detoxifying mechanism could exist in *B. pumilus*. In the case of MS32, there is a MFS transporter encoded directly upstream of *bacA*, the first biosynthetic gene of the cluster, perhaps this transporter takes over the function of the *bacE* permease. A similar BGC for bacilysin was also reported in *B. pumilus* 64-1, a strain with antimicrobial activity against pathogenic and drug resistant Gram-positive and Gram-negative bacteria. [105] Moreover, there is evidence in *B. subtilis* that bacilysin has a pleiotropic role as a signaling molecule, impacting protein expression levels and linked to late growth stage processes such as sporulation [244].

While the BGCs regions 1 and 12 of MS32 showed similarity to known clusters for lichenysin and bacilysin, respectively; the region 2 is a candidate for production of a lanthipeptide of the class III 3.4. Lanthipeptides correspond to ribosomally synthesized and post-translationally modified peptides (RiPPs). There are five types of lanthipeptides, all presenting a (methyl)-lanthionine ring as a distinctive feature, which also confers them stability. Lanthipeptides are classified according to the biosynthetic enzymes. A wide range of bioactivities have been associated to lanthipeptides, such as antimicrobial, antifungal, antiviral and antinociceptive [14, 354, 352]. Classes I and II are the most studied ones, while III-V are the least characterized with only a handful of known compounds [122], meaning that there is a great unexplored potential around these products.

"Secondary metabolites have multiple functions [73]:

- Competitive weapons used against other bacteria, fungi, amoebae, plants, insects, and large animals.
- Metal transporting agents.
- Agents of symbiosis between microbes and plants, nematodes, insects, and higher animals.
- Sexual hormones.
- Differentiation effectors."

Class III lanthipetides synthesis is mediated by a multifunctional LanKC enzyme, which carries out dehydration and cyclization steps. *B. pumilus* MS32 presents the characteristic LanKC, a serine protease, and an ABC transporter ATP-binding protein/permease in its BGC. Recently described, the andalusicin A represents a new family of class III lanthipeptides in *Firmicutes*, it was isolated from *B. thuringiensis* and presents antagonistic activity against other Gram-positive bacteria (*S. aureus*, *B. cereus* and *B. mycoides*) [122]. Interestingly, antibacterial functions have been associated mostly with class I and II lanthipeptides, and for class III absent or weak activity has been reported. Morphogenic functions have been described for class III lantipeptides in *Streptomyces*, but other functions such as antiallodynic, strong anti MRSA, antiviral activity against Herpes simplex, Dengue and Zika virus were also reported [135]. Moreover, lanthipeptides might have a signaling role in *Bacillus*, and despite their wide distribution and prominence, which points to a significant function in the genus, this BGC is understudied [124].

Remarkably, pumilacidin, a common lipopeptide among *B. pumilus* strains was absent from the predictions obtained for MS32 (Table 3.4). Similarly, it was reported that the type strain DSM27 is also deficient for production of this compound [233]. Activity against the phytopathogens *R. solani*, *P. aphanidermatum* and *S. rolfsii* was reported for the pumilacidin produced by the endophytic *B. pumilus* MAIIIM4a isolated from cassava collected in Brazil [223]. On the other hand, pumilacidin has also been linked as a probable cause to a small food poisoning incident due to contaminated rice that was improperly stored [107].

In a large scale bioinformatic study on 1566 genomes from *Bacillus* species, it was found that by far the most abundant BGC was the one dedicated to the synthesis of bacillibactin [124]. This catecholic siderophore is a nonribosomal peptide that constitutes the main iron scavenger for many *Bacillus*, it binds Fe 3+ with high affinity

[224, 257]. It was shown for *B. subtilis* that bacillibactin is part of the Fur (ferric uptake regulator) regulon, responsible for iron homeostasis [257]. Besides bacillibactin, an additional BGC for siderophore synthesis was identified in MS32 3.4, this region was also reported in *B. pumilus* SF-4 [149].

Plantazolicin (PZN) was first described in *B. velezensis* FZB42 (previously classified as *B. amyloliquefaciens* [79, 288]). Plantazolicin is ribosomally synthesized and post-translationally modified, resulting in a highly condensed Thiazole/Ozazole Modified Microcin (TOMM) with remarkable antibiotic specificity against *B. anthracis*, the causative agent of anthrax [228, 227]. Experiments based on crude cell extracts suggested nematocidal activity for PZN [200], however, no significant effect was found with purified PZN [227]. PZN production has been reported for some *B. pumilus* strains surface extracts, including DSM27, and for *B. subtilis* DSM32873 [233]. The BGC for synthesis of PZN was identified in *B. pumilus* MS32 and shows 100% similarity to that of DSM27 (MIBiG accession:BGC0001173), therefore it is likely that MS32 could be able to produce plantazolicin.

Regarding novel BGCs, table 3.17 showed that the third BGC predicted by antiSMASH in *B. pumilus* MS32 (Table 3.4) has remarkably high transcriptional activity at the transition point, and remains high at 7 hours. These genes appear as highly targeted for regulation by candidate sRNAs. No much characterization is available regarding this BGC with the potential to produce a NRP+Polyketide type of product. Within antiSMASH candidate BGCs are compared with a database of known biosynthetic gene clusters, for this case, similarity to paenilamicin and zwittermicin A was reported, however it is rather small, 18% and 21% of the BGC genes show similarity. The high transcript abundance observed at the transition point points to possible differentiation effector functions for this BGC product rather than a product for competitiveness. Nevertheless, identification of corresponding product is required, furthermore, characterization could lead to potential biotechnological applications.

All together, evaluation of the biosynthetic potential of *B. pumilus* not only enables the discovery of promising secondary metabolites, but also of their corresponding synthesizing enzymes, which could be exploited in areas such as drug engineering.

Unique proteins in *B.pumilus* MS32. There were 61 proteins encoded by the MS32 genome with no homologous counterparts among the remaining *B. pumilus* strains (Table 3.5), 31 of those are annotated as "hypothetical protein" but within the remaining annotated ones there are some interesting features:

- Novel lanthipeptide. The BGC for a class III lanthipeptide predicted by antiSMASH [30] is unique to MS32. As previously mentioned, little is known about class III lanthipeptides and therefore, there is unexplored potential around these compounds. Moreover, within this cluster there is a serine protease which could also serve for biotechnological applications.
- An RM system. There is a specificity subunit S from a type I RM-system unique to MS32, accordingly, the gene is preceded by a methyltransferase and followed by an ATPase and the corresponding type I restriction enzyme. This

RM-system might be part of the strain-specific signature of MS32. Interestingly, there is a second restriction enzyme downstream these genes which is also unique to MS32, however it was not recognized by the analysis against the REBASE [275, 274] database but it matched the protein family model for restriction endonuclease of the PGAP [312] annotation pipeline (HMM ID: NF016362), the protein is rather short (239 aa), and is probably a partial match (which also highlights the relevance of manual evaluation and curation of automatically generated annotations).

- Potential toxin. Similarly, the annotation for this sequence is based on a PGAP [312] model (HMM ID:NF015288) describing ETX/MTX2 family pore-forming toxins. Members of this group include mosquitocidal proteins usually found in *B. thuringiensis* and *Lysinibacillus sphaericus*. A more detailed look utilizing IDOPS [75] (See Chapter 6), a tool with high-quality and manually curated models for bacterial pesticidal toxins, revealed that this protein is not likely a true member of the MPP family (Beta pore-forming pesticidal proteins [65]). Nevertheless, the analysis of this protein with InterProScan [158] shows a match for Aerolysin-like toxin (ID:IPR004991 / PFAM 3318), a known member of this family is the Clostridium epsilon toxin ETX [54]. Even though the protein is not a pesticidal protein, the InterProScan match does not rule out potential pore-forming activity and therefore, it represents an optimization target to ensure safety for industrial applications. Additionally, the gene coding for this protein is part of a cluster of 6 proteins which seem unique to MS32, however, the surrounding genes only encode hypothetical or uncharacterized proteins, which leaves the open questions about its function and acquisition by *B. pumilus* MS32.

- Prophage related proteins. A type II toxin-antitoxin system PemK/MazF family toxin, a XRE family transcriptional regulator and a hypothetical protein are encoded by a region predicted as prophage by PhiSpy [4].

- Transposition and genetic mobility. Within the set of proteins unique to MS32, 4 sequences relate to genetic exchange. 1) DGQHR domain-containing protein, which is uncharacterized but the InterPro signature (ID:IPR017601) describes proteins with this match occur in contexts that suggest extensive lateral gene transfer. 2) TniQ family protein, similarly, tni genes seem involved in dissemination of integrons (Interpro ID IPR009492). 3) DDE-type integrase and 4) Recombinase family protein. Highlighting again that MS32 has a dynamic genome.

- RNA related. A DEAD/DEAH box helicase is predicted to be unique in MS32, proteins with this domain are involved in RNA metabolism, with roles such as modulation of RNA structures and transcriptional regulation. In *B. subtilis* the role of four DEAD-box helicases (CshA, CshB, DeaD, and YfmL) was studied, the authors found that different helicases have distinct roles in the physiology of the bacteria, with ribosome biogenesis and RNA degration being the major tasks [190]. In total, MS32 encodes 8 DEAD/DEAH box helicases.

- Carbon metabolism. MS32 encodes a mannonate dehydratase that mediates the reaction of D-mannonate to 2-dehydro-3-deoxy-D-gluconate + H2O, which can be directed to the penthose phosphate pathway [163]. This enzyme is part of the pentose and glucoronate interconversion pathway (KEGG orthology K01686). The protein is encoded by *uxuA*, which in *B. subtilis* is part of an operon involved in D-fructuronate degradation (Subtiwiki [250]). However, *MS32* seems to lack the accompanying fructuronate reductase encoded by *uxuB*, meaning that this catabolic pathway is incomplete. Instead, *uxuA* in MS32 is co-located with elements of the PTS sugar transporter system.

- An hydrolase. Also particular to MS32 is a MBL fold metallo-hydrolase. Proteins presenting the MBL fold structure constitute a superfamily of enzymes with great diversity of sequences and functions, with more than 81700 members registered at Pfam [23] (ID:PF00753). Enzymes of this superfamily tend to be promiscuous, participating in average of 1.5 reactions additionally to their native one [19]. DNA repair, RNA processing, detoxification, quorum-quenching, binding and transport are some of the functions elicited by members of this superfamily [19, 234]. The genome of *B. pumilus* MS32 encodes in total 10 proteins annotated as MBL fold metallo-hydrolases.

A recent comparative genomics study focused on the *B. pumilus* group (which includes the species *B. safensis*, *B. altitudinis* and *B. pumilus*), pointed out that the species *B. pumilus* exhibits the highest amount of non-core and strain-specific genes, which relates to the wide distribution and high genetic diversity of *B. pumilus* [108].

Comparison of *B. pumilus* with *B. subtilis* and *B. licheniformis*

Broadening the scope, it is relevant to compare *B. pumilus* also against other *Bacillus* species, specially those which are more extensively characterized and already established as microbial cell factories. Such insight allows to recognize features relevant for productivity, optimization targets, and differentiating aspects that would facilitate the implementation of *B. pumilus* as a productive platform in setups where other species are not optimal. Strain optimization based on knowledge gained from a pre-existing production host as a guide is a strategy themed Production Strain Blueprinting (PSB) [184]. PBS was previously applied in the development of another *B. pumilus* using *B. licheniformis* as reference, and resulted in promising protease production without complex process modifications [184].

General genomic features. Table 3.1 presented a characterization of *B. pumilus*, *B. subtilis* and *B. licheniformis* genomes. *B. pumilus* were smaller and described with a lower GC% than the other species. Genome size relates to complex aspects of bacteria such as generation time, replication rate, as well as physical space within a cell and energy supply [28]. Moreover, less redundancy is expected in smaller genomes, which in turn allows a more straightforward engineering and control of desired metabolic features [284].

Larger genomes are associated with bacteria occupying more complex and variable environments (such as soil or rizosphere), since they offer a bigger repertoire of genes to use as part of adaptation responses [366]. On the other hand, it is generally accepted that reduced and minimal genomes belong to bacteria limited to more stable niches, such as a host organism [28]. However, smaller genomes in free living bacteria also offer a competitive advantage in terms of energy saving and reproductive efficiency, as faster growth rates could be achieved when less resources are devoted to DNA, RNA, and protein synthesis and maintenance [182, 216, 28]. Following this idea, limited amounts of DNA can be stored in smaller cells, accordingly, since early descriptions it has been recognized that cells of *B. pumilus* tend to be smaller than those of *B. subtilis* [299], studies reported a mean cell width of 0,7 µm for *B. pumilus* and 0,8 µm for *B. licheniformis* and *B. subtilis* [202]. Additionally, it has been observed that the percentage of regulatory genes seems to increase with genome size, as gene expression must be regulated according to the available energy supply [28, 284]. Therefore, smaller genomes observed in *B. pumilus* could facilitate a more straight forward engineering by having less redundant elements [284].

The GC content of a given organism, and even a microbial community, is affected by external environmental factors [93]. Moreover, it can be associated to the lifestyle and the energetic context of a microorganism, as well as to the genetic and biochemical properties of genomes [28, 212]. Lower GC percentages relate to less requirement of phosphorus and nitrogen for DNA synthesis, additionally GTP and CTP nucleotides are energetically more expensive than ATP and UTP, therefore synthesis costs can be saved in organisms with lower GC content [216, 93, 28]. The lower GC content observed in *B. pumilus* (Table 3.1) genomes could confer the species this kind of advantages.

The number of rRNA copies correlates to the number of tRNA genes in order to support efficient protein synthesis [162]. The tRNA translation machinery in fast growing bacteria is better optimized to codon usage than in slow-growing ones, and duplication of tRNA genes allows to increase transcript amounts for rapid growth and replication [343]. Rapid growth and efficient protein synthesis are, of course, desirable features in productive strains.

The amount of copies of rRNA operons is related to bacterial adaptations to cope with fluctuating resource availability [57, 174, 277, 357]. A study analyzed soil bacteria and their response when exposed to highly nutritious medium. The bacteria able to adapt faster and rapidly form colonies carried in average 5.5 copies, while in contrast, the genomes of slow responding bacteria contained an average of 1.4. [174]. Generally, bacteria with more operon copies have a higher supply of ribosomes, which in turn allows them to support better growth rates and adapt to changing environmental conditions [174, 357]. More recently, a study proposed multiplicity of rRNA operons as a strategy to ensure genome stability by preventing over-saturation of these operons by RNA polymerases. Such saturation induces DNA replication blockage and breakage events that could be lethal to the cell [92]. For *B. subtilis*, it was shown that varying numbers of rRNA operon copies impacted not only growth rates, but also sporulation frequency, competence development, and motility processes [357]. *B. pumilus* genomes presented more rRNA operons than *B. licheniformis* and less than *B. subtilis*, pointing to good adaptation capabilities regarding protein synthesis and growth.

Further particular features of *B. pumilus* MS32. Among the elements unique to *B. pumilus* MS32, proteins related to carotenoid biosynthesis pathway were identified. Carotenoids in bacteria function as photoprotective agents, conferring resistance against damage caused by UV radiation [226]. This protection is not only important for the vegetative cell, but also for the spores, as it contributes to their endurance of environmental conditions. A wide range of pigmentation such as pink, red, yellow and orange have been reported in spore forming colonies of several *Bacillus* species [171].

Melanins and carotenoids are common photoprotective pigments. For example a melanin-like compound is found in the coat layer of *B. subtilis* spores as protection against solar radiation [272]. Pigment production is related to growth conditions. For example, colonies of the marine isolate *B. pumilus* SF214, showed a strong orange-red pigmentation when grown at 25 °C in contrast to the white colonies observed at 42 °C [171]. Another study on the same isolate observed that high pigment synthesis is under strict regulation and takes place during late stages of stationary growth [213]. Interestingly, the authors observed that only a portion of the cell population was able to generate the pigment, and that proportion increased at late stationary phase. However, pigment production and sporulation appear to be mutually exclusive developmental fates for a given cell. In the case of *B. pumilus* SF214, pigment production confers resistance against oxidative stress for vegetative cells and spore protection could be attributed to pigment-independent mechanisms [213]. Perhaps carotenoid compounds also benefit *B. pumilus* MS32 regarding stress endurance.

Another unique feature of *B. pumilus* MS32 is a protein annotated as "carbon-nitrogen hydrolase family protein", KofamScan [13] assigned the identifiers K01501: nitrilase [EC:3.5.5.1] and K18282: cyanide dihydratase [EC:3.5.5.-] to this sequence. Interestingly, this cyanide dehydrogenase, which catalyzes the reaction of Hydrogen cyanide + 2 H2O <=> Ammonia + Formate, has been reported only in few bacteria, with the most characterized ones described in *B. pumilus* and *Pseudomonas stutzeri* [66]. It was observed that *cynD*, the gene coding for this enzyme, is associated exclusively to some land isolates of *B. pumilus*, while it is absent from strains of marine environments [108]. Cyanide degrading nitrilases from *B. pumilus* are of interest for detoxification of industrial wastewaters contaminated with cyanide. An advantage of this type of enzymes is that they do not require cofactors or secondary substrates, therefore, research is conducted to improve the catalytic properties of CynD from *B. pumilus* [66].

Biosynthetic Gene Clusters. The analysis done with antiSMASH [30] predicted 13 BGCs for *B. pumilus* MS32, 14 for *B. subtilis* and 11 for *B. licheniformis*, common products included: bacillibactin and fengycin. Bacilysin is shared with *B. subtilis* and lichenysin with *B. licheniformis* (Table 3.4). The abundance and types of BGCs can also reveal insights about interactions between species and their evolutionary relationships. According to a recent study profiling BGCs in 4268 *Bacillus* genomes [350], distribution of BGCs is correlated with phylogeny in *Bacillus*. Closely related species present more similar BGCs than distantly related ones. This observation has also consequences for antagonistic interactions, as BGCs usually encode some sort of resistance mechanism to protect the host cell, and therefore bacteria with similar BGCs would exhibit lower inhibition between each other. In microbial communities where these bacteria co-exist, they could adapt a cooperation strategy in which common goods are shared by closely related species, while more distant ones are antagonized by the products of the BGCs [350]. The authors found that in average 11.6 BGCs could be identified per genome. Interestingly members of the *B. subtilis* clade, presented the most BGCs, 13.1 per genome (other clades such as *B. cereus* and *B. megaterium* had 11.7 and 7.4 BGCs per genome, respectively), which could be an adaptation to competitive environments such as plant rizosphere. Moreover, the study revealed a positive correlation between antagonism and phylogeny, specially in antagonistic strains with abundant BGCs, for example, *B. pumilus* ACCC04450 showed weak antagonism against *B. amyloliquefaciens* ACCC19745, as both belong to the *subtilis* clade.

Signal Peptides. Notably, a study in 2004 characterizing the genome of *B. licheniformis* DSM13 found 689 proteins with predicted signal peptides by SignalP [271]. During earlier developments of this kind of software, accurate discrimination between real signal peptides and N-terminal transmembrane helices was problematic, therefore false positive predictions were abundant, and complementary analysis with tools like TMHMM were necessary to filter the results. This issue was tackled since the release of SignalP version 4.0, which implements a neural network based approach trained to discern sequences with transmembrane regions [254]. Comparable amounts relative to proteome size were identified for *B. pumilus* MS32, *B. licheniformis* DSM13 and *B. subtilus* 168 (Table 3.9), with *B. pumilus* depicting a slightly higher percentage.

Identification and characterization of the set of signal peptides of a production host is relevant for optimization of heterologous protein secretion. Libraries of different signal peptides accompanying a target protein are screened in search for optimal secretion partners [106, 77]. A study on a large library of signal peptides (173 from *B. subtilis* and 220 from *B. licheniformis*) for production of protease BPN' from *B. amyloliquefaciens* in three expression hosts (*B. subtilis* TEB103, *B. licheniformis* strains DSM13 and H402), observed a similar relative performance of the majority of signal peptides in the three *Bacillus* [72], pointing to some secretion conserved properties shared by these organisms. Prediction of the optimal "signal peptide - target protein" combination is almost impossible to achieve (currently), nevertheless, some insight could be gained by transferring the knowledge of a certain host to a closely related one, at least during first rounds of screening.

Functional annotation. The assignment of COG categories identified a higher amount of proteins related to translation, transcription and replication functions in *B. pumilus* MS32 than in *B. licheniformis* DSM13 and *B. subtilis* 168 (Table 3.7). This could be indicative of higher metabolic capacity to sustain (faster) growth and reproductive functions in *B. pumilus*. Similarly, signal transduction associated proteins were slightly higher for *B. pumilus*. Signal transduction systems allow the bacteria to sense changes in environmental and intracellular conditions, so adaptive metabolic, behavior and/or physiological responses are triggered [110]. Improved capacity to sense and adjust to (perhaps a wider/more specific) set of signals could give *B. pumilus* an self fine-tuning advantage when facing challenging conditions, for example the diverse stress sources within a bioreactor (See Chapter 1).

There were also differences identified regarding the COG category of "P:Inorganic ion transport and metabolism" (Table 3.7). Notably, within this category, transport systems dedicated to iron acquisition were highly abundant in the *B. pumilus* MS32 protein set with no homologous counterpart in the other species. Out of 20 sequences, 12 were related to iron transport, mostly putative ABC-type systems associated to siderophore mobilization. Interestingly, there was also a cluster of three consecutive proteins annotated as "Heme-binding NEAT domain protein". The NEAT (NEAr-iron Transporter) domain has been previously associated to pathogenic bacteria, as it facilitates the acquisition of heme-iron from host hemoglobin during infection [10, 142]. However, more recent research revealed the NEAT domain distributed within *Firmicutes* and also present in non-pathogenic species associated with soil and plant environments, such as *Paenibacillus polymyxa* and other strains of *B. pumilus*

[142].

Moreover, by complementing these results with transcriptomic data, it is evidenced that dedicated Iron transport systems are not only more abundant in *B. pumilus* MS32 than in *B. subtilis* 168 and *B. licheniformis* DSM13 genomes, but these elements are also highly transcribed, particularly at 4 and 7 hours of the fermentation (Table 3.18).

Iron is essential to support growth, bacteria evolved high-affinity iron uptake strategies, such as siderophores and ABC-type transporters with specific surface receptors to facilitate the incorporation of complex and non-complex iron sources [134, 38]. Limited iron availability has great impact over central carbon and nitrogen metabolism, as it is required as a cofactor for many proteins (metalloenzymes, respiratory proteins and cytochromes for example). In *B. subtilis* the TCA (tricarboxylic acid cycle) was observed to be significantly repressed during iron limitation, which has broader consequences since many of its intermediate products are also precursors for several other metabolites [298]. In many natural environments, low solubility and bioavailability of the Fe3+ ion make of iron a limited resource, diverse acquisition mechanisms benefit bacterial fitness and adaptation capacity [134]. For *B. pumilus*, its abundant iron transport systems (NEAT proteins, siderophores, ABC transporters) could represent a competitive advantage in such environments, particularly the NEAT proteins, which offer potential access to additional iron sources that might not be accessible for *B. subtilis* or *B. licheniformis*.

Another noteworthy difference was observed regarding transport and metabolism of nutrients between the studied *Bacillus* species (Table 3.7). The higher percentage of proteins related to "E: Amino acid transport and metabolism" in *B. pumilus* MS32 could be related to inter-species differences in substrate utilization capacities. This is useful when considering initiatives aimed to minimize and make use of substrates otherwise regarded as waste, such as agro-food by-products (peels, soybean residues, sugarcane bagasse and wheat bran for example) [207, 313]. An advantage for *B. pumilus* might come by superior importing of amino acids from the extracellular environment, consequently reducing the need for amino acid synthesis proteins which represent a higher metabolic cost for the bacteria [218].

The COG analysis suggests that *B. pumilus* might prefer and (given optimized conditions) outperform in presence of proteinaceous rich substrates, which together with the identified protease content and secretory machinery, profiles *B. pumilus* as a good candidate with potential for biotechnological application processes.

Comparative genomic analysis, such as those here presented, benefit from the study of which of those features are active and interacting at a given condition of interest, for example a productive process. The environment within a bioreactor is drastically different from a lab-scale culture. Consequently, it has been long recognized that most bacterial strains exhibit different performances between such conditions [342]. Even though multiple studies characterize *Bacillus* species from perspectives such as transcriptomics, proteomics and metabolomics, many of them are performed at laboratory scales. Therefore, some findings are not directly applicable for biotechnological engineering directed to optimize strains of industrial interest [125]. To overcome this gap, a set of small-scale fermentations were performed on *B. pumilus* MS32, *B. licheniformis* MW3 Δ *yqfD* and *B. subtilis* Δ *yqfD*.

4.3 Optimized RNA isolation protocol

RNA samples were obtained from small scale fermentations in order to conduct comparative transcriptomic analysis to further explore the potential of *B, pumilus* to be developed as a microbial cell factory. A critical step in such investigations is the adequate processing of the samples in order to purify RNA of enough quality for RNA-seq applications. This is necessary for reproducible results and achievement of biologically significant conclusions [287]. However, purification of RNA of high quality is often not a trivial task. Common challenges include: incomplete cell lysis, poor RNA precipitation efficiency, isolation out of complex media, and highly active RNases, which are a threat to RNA integrity. These challenges are of notable consideration when working with non-domesticated strains. A RNA isolation protocol optimized for *Bacillus* samples of industrial relevance was developed for this project. Such method was highly needed since standard approaches failed to deliver RNA of high quality from the samples of interest. The improved protocol overcomes common challenges of RNA isolation and presents modifications resulting in higher RNA yield, purity and integrity. The optimized protocol and the rationale behind it was prepared as a manuscript and submitted for publication (Chapter 6).

4.4 Observations from RNA-seq analysis

Once high quality RNA was purified and libraries sequenced, the comparative transcriptomic analysis contrasting *B. pumilus* and *B. licheniformis* species could take place. According to the literature reviewed for this work, this appears to be the first report of the tools Annogesic [363] and DP_GP_cluster [220] implemented for the analysis of *B. licheniformis* and *B. pumilus* transcriptomes.

Transcriptome sequencing. High abundance of rRNA derived reads is a common challenge in RNA-seq studies [327]. The rRNA can constitute more than 85% of the total RNA present in a prokaryotic sample [255, 68]. Without depletion or selective procedures, most of the reads would map to rRNA, which hinders the detection and study of other RNA species, such as mRNA and sRNAs, which are the often the focus of investigations. For eukaryotic samples, which have polyadenylated mRNAs, selection methods for polyA transcripts using oligo (dT) primers are available [327, 255]. However, most bacterial transcripts lack polyA tails, therefore, selective enrichment is not an option.

Subtractive hybridization has become the method of choice for bacterial rRNA depletion [255, 68]. This is the approach of kits such as the Ribo-Zero, which uses biotinylated rRNA capture probes which hybridize rRNAs and are removed from the sample using magnetic beads. This kit was reported to outperform the Ambion MICROBExpress™ Bacterial mRNA Enrichment and the Life Technologies RiboMinus Transcriptome Isolation kits, particularly impacting the detection of low abundant ncRNAs [255]. The Ribo-Zero Plus kit was used for this study, the succesfully rRNA depleted samples confirm the kit is suitable for the *Bacillus* of interest (Table 3.11), and the failed depletions could be attributed to other factors.

It is also relevant to report the TIN (Transcript Integrity Number) values, as RNA integrity is crucial for successful analysis of RNA-seq data. Calculation of TIN scores is recommended for quality assessment of RNA-seq data, moreover, since it can be determined at individual transcript level, TIN values are useful to correct biases arising from differentially degraded transcripts [300] [336]. The obtained TIN (Table 3.11) correspond to high-quality transcripts and reflect the previously generated RIN (RNA Integrity Number) values, which together indicate the successful implementation of the optimized RNA isolation protocol.

Predicted sRNAs. Another transcriptomic study on *B. pumilus* SCU11 (recently reassigned as *B. altitudinis* [85, 198]), found 84 putative sRNAs with sizes between 50 to 1058 nucleotides [353]. The custom-made sRNA database for *B. pumilus* used by Annogesic [363] included sRNAs from the SCU11 publication. The Blast+ search against the sRNA database step within Annogesic [363] found 7 sRNAs with homology to previously reported sRNAs in *B. altitudinis* SCU11, which point to conserved sRNAs within closely related *Bacillus*. Those include: Bpsr1, Bpsr34, Bpsr70, Bpsr92, Bpsr139, Bpsr178 and Bpsr193. When Bpsr139 was deleted from the SCU11 strain, cell growth decreased in M9 medium but not in LB [353]. Bpsr34 also matched the RFAM entry RF00168, which corresponds a Lysine riboswitch, a sensor for lysine that modulates expression of genes involved in lysine biosynthesis, transport and catabolism [159, 121].

Three more candidate sRNAs of *B. pumilus* MS32 predicted by Annogesic [363] were found to match RFAM [121] entries (Table 3.14). The predicted sRNA1 matched RFAM:RF00379, which describes a cyclic di-AMP riboswitch (previously known as YdaO/YuA leader). In *B. subtilis* there are two instances of this riboswitch, one associated with *kimA* and another with the *ktrA-ktrB* operon [250], these genes encode high affinity potassium transporters which are under sophisticated control mechanisms in order to keep potassium homeostasis [305, 126]. Potassium is essential for growth, pH maintenance, ribosomal and enzymatic functions [126]. The riboswitch has been associated to sporulation, osmotic stress and cell wall metabolism [268].

According to the Proteinortho [188] analysis, a KtrA homolog protein was not found in *B. pumilus* MS32, but it is present in *B. licheniformis* MW3, while KimA is present in the three *Bacillus* species. Moreover, for MS32 the genomic location of the candidate sRNA1 is not upstream KimA, like in *B. subtilis*, but rather it is found in association with a dicarboxylate/amino acid:cation symporter. According to the Transporter Classification Database (TCDB) [281], this transporter belongs to the DAACS family, which "catalize the Na+ and/or H+ symport together with (a) a Krebs cycle dicarboxylate (malate, succinate, or fumarate), (b) a dicarboxylic amino acid (glutamate or aspartate), (c) a small, semipolar, neutral amino acid (Ala, Ser, Cys, Thr), (d) both neutral and acidic amino acids or (e) most zwitterionic and dibasic amino acids" [281].

The remaining RFAM [121] matches within the Annogesic [363] prediction for *B. pumilus* MS32 corresponded to candidates sRNA30 and sRNA31, both matching RFAM:RF00023 describing Transfer-messenger RNA (tmRNA), also called SsrA. SsrA is also found in *B. subtilis* and *B. licheniformis* (Table 3.14). This RNA has functional and structural properties from both tRNA and mRNA, it is highly versatile and has a key role in ribosomal recycling by rescuing it from stalled processing of defective mRNAs (e.g. those without/with poorly efficient stop codon or when the corresponding tRNA is scarce)[323]. SsrA adds a peptide tag to the abnormal product targeting it for proteolytic degradation [168, 153]. This quality control function is relevant for production platforms. Often protease-negative strains are used as industrial production hosts; but without the corresponding protease degrading the SsrA-tagged molecules, the product of interest might become contaminated with nonfunctional proteins [179]. In *B. subtilis* a reduction in amylase yield was reported in knockout *ssrA* mutants [113]. SsrA has also been described as necessary for efficient growth under stress conditions and for spore formation [1].

For *B. pumilus* MS32 the candidates sRNA30 and sRNA31, predicted by Annogesic [363], were both identified as SsrA (RFAM:RF00023), they are 261 and 67 nt, respectively and separated by 10 basepairs. Most bacterial tmRNA molecules have a size between 325 and 400 nucleotides [153]. It is possible that sRNA30 and sRNA31 actually correspond to a single RNA and were predicted as two due to its coverage pattern.

From the putative sRNAs detected in *B. pumilus* MS32 by Annogesic [363], the top three transcriptional activities were observed for Bpsr139, Bpsr193 and Bpsr92. Bpsr139 belongs to the cluster 14 (Figure 3.10), its transcription was downshifted at 2.5 and 4h, then increased to a maximum TPM (replicate mean) of 2741.7 at 19 hours. Members of cluster 14 also include SlrA, CtpB and CspC. Bpsr193 and Bpsr92 follow the pattern described in cluster 17 (Figure 3.11), with maximum TPM values of 951.7 and 1650.3 at 7h for Bpsr92 and Bpsr193, respectively. Interestingly, key regulators such as KinA, SinR, SinI, SigB and SwrA are also found within this cluster (Table 3.23).

Regarding the scan against the RFAM [121] database of the genomes of interest, Table 3.14 suggests that several of the RNAs known in *B. subtilis* (and functionally characterized) are also present in *B. pumilus* and *B. licheniformis*, which points to shared regulatory mechanisms. Examples include FsrA, BsrC, BsrG, SR1, RoxS and several riboswitches; further research could confirm if they play similar regulatory roles within these related *Bacillus* species.

BsrF (RFAM:RF01411) seems to be present in *B. pumilus* and *B. subtilis* but not in *B. licheniformis* (Table 3.14). Research in *B. subtilis* characterized BsrF as a probably noncoding sRNA located in the intergenic region between *yobO* and *csaA*, and its terminator region overlaps by 30 bp with that of *csaA* [260]. CsaA has been described as a secretion dedicated chaperone, which was significantly induced in response to secretory stress in *B. subtilis* [318, 148]. Expression of *bsrF* was reported in *B. subtilis* at all growth phases and decreased during sporulation, the authors proposed BsrF as implicated in fine-tuning of gene expression since major growth defect phenotypes were not observed in deletion or over expression mutants [260], the study searched for BsrF homologues in other Gram-positive bacteria and only *B. amyloliquefaciens* presented a highly homologous sequence. The search using RFAM covariance models suggests that BsrF is also distributed within *B. pumilus*, as it was detected not only for MS32, but also in all the other *B. pumilus* strains analyzed.

Table 3.14 also depicts a total of 19 matches for RFAM:RF01458 identified in *B. pumilus*, this RNA was not predicted in *B. subtilis* or *B. licheniformis*. The RFAM entry corresponds to rli23, a sRNA initially described in the Gram-positive *Listeria monocytogenes*, were it is located antisense to the transposase gene *lmo0172* [289]. For *B. pumilus* MS32, every instance of this RNA was located in the opposite strand of an IS3 family transposase. Therefore, this sRNA candidate could be involved in mobile genetic element regulation. According to the RFAM database [121], the rli23 model also matches sequences found in other Gram-positives, including: *B. amyloliquefaciens* FZB42, *Lactobacillus dextrinicus* DSM20335, *Lysinibacillus sphaericus* OT4b.31 and *Bacillus nakamurai*.

The stringent filtering parameters applied, such as the cutoff for the secondary folding energy change and the condition of being found in all replicates helped to keep only high-quality candidates with the drawback of reduced total sRNAs identified. The Annogesic settings used the default size limits of 30 to 500 nt for the prediction, therefore, it is possible that longer sRNAs were missed. A limitation for the prediction of sRNAs is the lack of dRNA-Seq data. Annogesic has a functionality to process dRNA-Seq data (TEX +/-) to produce TSS (Transcript Start Site) predictions and that information is used to identify UTR-derived sRNAs. By complementing the conventional dataset with dRNA-Seq, the detection could be improved.

Antisense activity of predicted sRNAs. For *B. pumilus* MS32 the candidate sRNA Bpsr193/srna0 was found encoded in the complementary strand of an uncharacterized protein presenting a DUF348 domain and elicited high transcriptional activity (Table 3.15). Its transcription profile matched that of cluster 17 (Figure 3.11). It was proposed that a protein family in firmicutes containing the DUF348 signature could be functionally equivalent to the Resuscitation-Promoting Factors of actinobacteria, which play a role in cell wall modifications that take place during restoration of active growth in dormant cells [265].

In *B. pumilus* MS32 and *B. licheniformis* MW3 sRNAs candidates were found opposite to a "NCS2 family permease" encoding gene (Tables 3.15, 3.16), which in *B. subtilis* corresponds to a hypoxan-thineguanine permease PbuO. In *B. subtilis* this protein is repressed in the presence of purine nucleotides and it is regulated by PurR [250]. The difference in transcriptional activity of the candidate sRNAs targeting this mRNA could implicate that the potential regulatory function is required only during the transition point, but is perhaps stronger for *B. licheniformis*.

This study presented a catalog of candidate regulatory anstisense sRNAs for *B. pumilus* MS32 and *B. licheniformis* MW3. The interaction of those sRNAs presented in Tables 3.15 and 3.16 with their putative antisense targets remain to be experimentally confirmed. One advantage of these prediction is that they are based on transcriptomic data rather than just genomic sequences, which allows the identification of transcriptionally active elements. This is relevant to generate a narrower list of candidates for further experimental validation.

Clustering of transcriptional profiles. An useful way to reveal regulatory mechanisms behind response and adaptation to environmental changes is to summarize transcriptome-wide data into groups of features with similar transcriptional dynamics across a time series [220]. Genomic features with similar transcriptional trajectories tend to share biological functions [83]. This approach is also helpful in the characterization of genes with unknown functions, as association with a particular set of functionally characterized genes gives an indication of potential annotations [220]. The software DP_GP_cluster was selected for this task and returned 18 clusters for *B. pumilus* MS32 and 25 for *B. licheniformis* MW3 ∆ *yqfD* (Figures 3.9, 3.10, 3.11), 3.12, 3.13, 3.14 and 3.15).

To prepare the data for DP_GP_cluster, mean TPM values were normalized using the hyperbolic arcsine transformation (asinh). Other approaches to stabilize the variance across mean values in RNA-seq data analysis include normalization by log transforming gene counts. The issue of zero values is circumvented by adding a pseudocount, for example of 1, to every gene count. However, there is a detrimental effect in proceeding like this, as low counts are disproportionately increased in comparison to genes with higher gene counts [156]. By using the asinh function the need for a pseudocount is eliminated, as the function deals with zero values and has a similar normalizing effect as the natural log function. Asinh transformation has been described as outperforming other methods and is recommended as the transformation of choice for coexpression analysis [156].

These kind of analysis open the door for further investigations, for example, by searching common sequence motifs within a cluster of interest, recognition sequences of key regulators can be identified and regulatory networks behind specific cluster trajectories revealed.

Differential transcriptional activity analysis. According to the DEG analysis implemented within READemption [104] the top most significantly regulated genes in *B. pumilus* MS32 at the transition point of the fermentation were those encoding for: oligoendopeptidase F, WD40 repeat domain-containing protein, and a stress protein.

The Proteinortho [188] analysis revealed that the oligoendopeptidase F is homologous to a protein encoded by *pepF* in *B. subtilis*. The PepF peptidase is located at the cytoplasm and has been investigated for its role in inhibition of sporulation. The PhrA peptide has been proposed as a target of this oliegoendopeptidase [161]. PhrA is part of the phosphorelay signal transduction system that orchestrates the initiation of sporulation in *B. subtilis* [155]. The PhrA pentapeptide has a specific inhibitory role over the phosphatase RapA, which in turn affects the phosphorylation levels of the response regulator Spo0F. Derepression of Rap phosphatases by Phr peptides leads to sporulation inhibition [161, 155].

This regulatory mechanism is essential for communication and coordination of the bacterial population in order to determine whether the cells remain in a vegetative growth state or if developmental differentiation processes such as those for growth, competence, and sporulation physiological states should take place. Therefore, it makes sense that this endopeptidase has a differential transcriptional activity at the transition phase of the culture, when many regulatory mechanisms are active and contribute to determine the cells fates. Moreover, a study of competent and non-competent subpopulations of *B. subtilis* found a higher abundance of PepF in the competent subpopulation [32]. As sporulation and cellular competence are mutually exclusive physiological states, it is reasonable to found this peptidase enriched in the competent subpopulation.

The endopeptidase was found to follow a similar transcriptional profile as reported for *B. subtilis*, where *pepF* is transcribed at a low level during the exponential growth phase and a two fold increase was detected at the transition to stationary phase [161]. For *B. pumilus* MS32 the expression profile was similar, showing a peak at the transition point and then down regulated during the stationary phase (cluster 16, Figure 3.10). For *B. licheniformis* it was found as member of the cluster 20 (3.14), which depicts a similar pattern until the transition point, but afterwards remains up regulated during the stationary phase.

Delay or inhibition of sporulation is an attractive target for optimization of industrial production platforms, since cells devoted to the production of spores do not contribute to the fermentation productivity. For *B. subtilis* there is a paralogue to PepF, called YusY. In *B. subtilis* JH642 it seems like *yusY* and the neighboring *yusX* correspond to a single ORF, which product is truncated and inactive [161]. However, for other *Bacillus*, mutation or reduced expression of YusX (the putative YusXY) or YusZ has been described as a method to optimize heterologous protein secretion. When deleted, secretion of heterologous AmyQ in *B. subtilis* was reported to increase over 200%, similarly, when the corresponding homologs in *B. licheniformis* were mutated, secretion of an heterologous protease also increased [240]. Even though there is still characterization pending for these proteins, it is clear that these endopeptidases have key functions impacting the yield of secreted proteins and offer a promising target for strain optimization.

Another significantly up-regulated feature in *B. pumilus* MS32 is a WD40 repeat domain containing protein. The transcriptional activity is low at 2.5h, increases at 4h, peaks at 7h and remains high until the last sampling point. Therefore, this protein could have an important role during the productive phase of the fermentation. Interestingly, there is no homologous WD40 repeat domain containing protein found in *B. subtilis* 168 or *B. licheniformis* MW3, according to the Proteinortho [188] analysis. The PGAP annotation of this protein in *B. pumilus* was derived from detected homology to the RefSeq entry WP_012009237. This protein is found in *B. pumilus* SAFR-032, a strain famous for its tolerance to environmental stresses [317]. Proteinortho [188] also revealed that this protein is present in every *B. pumilus* strain analyzed, meaning that this could be a protein particular to the *B. pumilus* group.

There is no much characterization available for this WD40 repeat domain containing protein. However, KofamScan [13] assigned the KEGG [163] entry K20332 to the protein. The KEGG entry is associated to Quorum sensing pathways (ko02024). In particular, K20332 has been characterized as part of the toxoflavin biosynthesis pathway in *Burkholderia glumae*. In eukaryotes, the WD40 proteins are abundant and usually function as scaffolds for the assembly of complexes with roles on signal transduction, transcriptional regulation, and ubiquitin-dependent protein degradation [145]. The prokaryotic counterparts have been less characterized and seem less abundant, nevertheless, involvement in signal transduction, protein folding and transcription has been suggested [145].

A stress protein was also found in the top most significantly up-regulated features of *B. pumilus* MS32. Homologous proteins are present in *B. licheniformis* and *B. subtilis* and are encoded by *yvgO*. No COG, KEGG or GO term assignment was found for this product. The YvgO protein in *B. subtilis* is secreted and has been described as a general stress protein associated to survival in ethanol stress conditions [250]. The expression of YvgO is under control of the alternative SigB sigma factor and was found to be induced under phosphate starvation conditions [259]. After the significant up-regulation of this protein at 4 hours, the transcriptional activity further increased at 7 and 19 hours, reaching a maximum of 80564 TPM at 7 hours. The transcriptional profile was assigned to Cluster 17 of figure 3.11, in which SigB is also found. This stress protein in *B. licheniformis* MW3 has a different transcriptional profile and did not reached such high TPM values. For *B. licheniformis* the maximum transcriptional activity of the stress protein happens at 19 hours reaching 475 TPM, while the previous sampling points presented TPM values below 55.

Proteases and protein secretion

Besides the biotechnological and commercial relevance of *Bacillus* proteases, these enzymes also play a key role in posttranslational regulation and maintaining protein homeostasis within the cell. Therefore, proteases of *B. pumilus* MS32 and *B. licheniformis* MW3 were further analyzed and compared.

The wall associated WprA was highly transcribed by *B. pumilus* MS32 (Table 3.19). WrpA degrades misfolded or slowly folding secretory proteins, making it a relevant quality control system. Without this function, defective proteins can accumulate and aggregate at the cell wall, which interferes with cell elongation and cell wall synthesis, this can result in cell lysis [132]. In a productive fermentation, proteases released by cell lysis can affect product yield, moreover, production of properly folded proteins with the desired specificity might become affected if this kind of quality control is absent. It was reported that WrpA enhanced pullulanase production and specificity in *B. subtilis* [369]. This quality control function might contribute as well in *B. pumilus* productive systems.

Regarding the Isp protease, the transcriptional profiles were different for *B. pumilus* and *B. licheniformis* (Tables 3.19 and 3.20). For *B. pumilus isp* was more actively transcribed in early time points and then a downshift was observed at 7 and 19 hours. While for *B. licheniformis* the maximal activity was detected at 2.5 hours followed by a downshift at the transition point and subsequent increase in transcription during stationary growth. A role in protein processing during stationary phase has been described for Isp, and it has not been linked to improved product production [132]. Expression of this intracellular protease is repressed during growth in presence of branched chain amino acids [132], which seems to be the case for *B. pumilus* MS32.

AprX protease is non essential for growth or sporulation, it is a member of the LexA regulon involved in the SOS response. Deletion of its encoding gene has been associated to reduced secreted product degradation by AprX released by cell lysis at late growth phase [132]. *aprX* showed high transcriptional activity at 19 hours for *B. pumilus* while for *B. licheniformis* the maximum TPM was 4.1. This could represent another optimization target.

MlpA activity has been reported to antagonize AprE expression, as it is involved in regulation of extracellular proteases without DegU mediation, potentially by degradation of a transcriptional regulator of *aprE* [132]. For *B. pumilus* maximal transcriptional activity of *mlpA* occurred at 2.5 hours and then it decreased as the fermentation advanced (Table 3.19). In the case of *B. licheniformis* the maximal activity also was observed at 2.5 hours, however a second increase was detected at 7 hours followed by a small decrease (Table 3.20). Membership to a particular regulon for this protease has not been determined [132], deeper examination of the clusters describing the transcriptional profile of the encoding gene could be associated to potential regulon.

Transcriptional profiles of *ftsH* differs between *B. pumilus* and *B. licheniformis* (Tables 3.19 and 3.20). While the highest TPM for *ftsH* for *B. pumilus* occurred at the transition phase (1949.4), in *B. licheniformis* a two-fold downshift was observed at this point, followed by a second activity peak at 7 hours. In *B. subtilis*, the membrane protease FtsH participates in sporulation, protein quality control, secretion, cell division, cell envelope stress and biofilm formation. The divergence in the observed transcriptional profiles suggests different requirements of this protease by these *Bacillus* species, perhaps according to the subpopulations defined during the fermentation and groth phase. The peak at 7 hours in *B. licheniformis* contrasting to the stable levels in *B. pumilus* could be related to initiation of sporulation processes, as known targets of FtsH are Spo0E and Spo0M, [250]. Alternatively, it could be related to competence development functions of FtsH, as it was observed in *B. subtilis* that FtsH is involved in competence, although the exact mechanisms is still to be elucidated, lost of competence was reported in *ftsH* mutants [27]. Moreover, it has been observed that the SigW regulon is induced in *ftsH* knockout strains of *Bacillus subtilis*, however the mechanism behind is unknown. It was postulated that FtsH could be involved in degradation of SigW or that FtsH absence leads to the accumulation of products which trigger the SigW regulon [365]. The high TPM values observed at early growth stages could be associated to the role of FtsH during cell division, as it has been reported that FtsH locates at the septum in exponentially growing *B. subtilis* [360]. In *E. coli*, FtsH has a role keeping balance between Sec translocase components by degrading excess of SecY that is not in complex with SecE, however this function has not been confirmed in *Bacillus*. Interestingly, FtsH has been proposed as a target to inhibit biofilm formation [360].

It is of notice the different transcription profiles of *htpX* encoding the membrane metalloprotease HtpX (Tables 3.19 and 3.20). While for both *Bacillus* the encoding gene reached maximal TPM at the transition point, for *B. pumilus* the transcriptional activity of this gene is high since the first sampling point and remains relatively high until a down shift at 19 hours. In contrast, for *B. licheniformis* a strong 3.9 fold up shift was observed a the transition point. followed by decreasing TPM levels. This protease is involved in membrane quality control and in response to stress [132], particularly that associated with growth at high temperatures [195]. In *B. subtilis* it has been proposed that FtsH and HtpX have partially overlapping functionalities, as growth under heat stress is impaired when both are absent but not if one of them is present [195]. Perhaps the up shift at the transition point in *B. licheniformis* means that this organism relies more on HtpX than in FtsH as response to the stress associated to the fermentation conditions.

Another difference between *B. pumilus* and *B. licheniformis* was found regarding the SppA serine protease(Tables 3.19 and 3.20). This protein presented maximal transcriptional activity at 4 hours in *B. pumilus* and remained relatively high during the rest of the fermentation. In contrast, for *B. licheniformis* this protease reached peak activity at 2.5 hours and remained down shifted until the end of the fermentation. SppA has a role for optimal translocation and processing of secretory proteins, as it cleaves signal peptide remnants and contributes to keep the membrane and secretory machinery clear. It was shown in *B. licheniformis* that heterologous production of nattokinase and α-amylase was increased by over expression of SppA [42]. The naturally higher and more stable levels of *sppA* transcripts detected in *B. pumilus*

MS32 during the fermentation could confer it an advantage over *B. licheniformis* regarding efficient secretory capacity. There are four candidate sRNAs predicted to interact with *sppA* mRNA in *B. pumilus* (Table 3.19) which could contribute to its transcriptional activity, for example, a strong positive correlation of 0.953 (pval=0.047) was found for the putative srna2 and *sppA* mRNA, Annogesic [363] predicted energies for this interaction to be -19.12, -10.49 and -16.449 as determined by RNAup, RNAplex and IntaRNA, respectively.

The protein secretion apparatus of *Bacillus* species has been subject of intensive study. Development of industrial production platforms often require optimization around bottlenecks that emerge due to the increased demand on the secretory machinery [240]. The potential regulatory sRNAs candidates proposed by the Annogesic [363] analysis could be further investigated, and if regulatory activity is confirmed, such insight could be used in fine-tuning and optimization of productive strains, which could help to overcome bottlenecks associated to the high demand on the secretory apparatus and secretion related stress in bacteria.

For example, in *B. subtilis* overexpression of the signal recognition particle *ftsY* has been associated to increased secretion of heterologous proteins [240]. The candidate regulatory sRNAs proposed by this study could be of aid regarding this matter, since for both *B. pumilus* and *B. licheniformis* the transcriptional activity of *ftsY* showed a decrease during the productive phase (Tables 3.21 and 3.22).

The twin-arginine (Tat) translocation pathway of *Bacillus* remains as a less characterized and exploited resource in comparison to the Sec pathway [77]. However it has tremendous potential, as it offers the capability to export fully folded proteins (even enzyme complexes and proteins associated with cofactors) [76, 258]. Tables 3.21 and 3.22 showed that maximal transcriptional activity of *tatC* and *tatAE* occurred at 7 hours for *B. pumilus* MS32 and at 4 hours for *B. licheniformis* MW3. Several potentially regulatory sRNAs were proposed as interacting partners for these mRNAs. The transcriptional profile in MW3 is similar to the previous report of the Tat system being more actively transcribed during the transition phase in *B. altitudinis* BA06. This study contributes to the characterization of the Tat pathway in biotechnologically relevant *Bacillus* species, which could aid to unlock its potential for industrial application purposes.

Flotillins are protein chaperones associated to the membrane and contribute to the organization of lipid rafts in eukaryotes. *B. subtilis* produces two flotillin-like proteins, FloA and FloT, which have been suggested to have similar roles as eukaryotic flotillins, being involved in the organization of membrane micro domains with functions related to signal transduction, transport, and protein secretion [34]. *floA* and *floT* were identified in *B. pumilus* MS32, and presented higher transcriptional activity in early stages of the fermentation (Table 3.21). While in *B. licheniformis* MW3 transcription for FloA peaked at the transition point and transcripts remained abundant during the rest of the fermentation. Interestingly, the Proteinortho [188] analysis, neither the PGAP based annotation, could identify an homologous protein to FloT in *B. licheniformis*. It might be the case that its sequence diverges and is below detection parameters of the software employed, given that it has been reported that *B. licheniformis* and other close relatives of *B. subtilis* encode both *floA* and *floT* [34]. In *B. subtilis* it has been reported that flotillins contribute to organize the membrane environment for the proper functionality of the Sec machinery and secretion functions were altered and reduced its absence [17].

4.4.1 Comparative analysis of key regulatory systems and potential interacting sRNAs

To further characterize *B. pumilus* and understand its potential in comparison with other established production platforms, it is useful to examine known key regulators that determine, regulate and control the gene expression behind major metabolic events. Some examples were presented in Tables 3.23 and 3.24 and are discussed below.

DegS-DegU

The two component system response regulator DegU-DegS has a key role in synthesis of degradative enzymes (such as aprE), competence development, biofilm formation and capsule biosynthesis [204, 194]. The sensor kinase DegS phosphorylates DegU. Phosphorylated DegU (DegU-P) negatively impacts competence development and swarming motility, while enhancing its own activity in a autoregulatory loop. DegU-P induces production of extracellular degradative enzymes and poly-γ-glutamic acid [204]. The small protein DegQ stimulates DegS activity over DegU and is required for complete activation of DegU [194].

In *B. pumilus degU* presented high transcriptional activity since the first measurement (1608.6 TPM), then peaked at the transition point with 2588.2 TPM and remained high until the end of the fermentation (2211.4 TPM at 19 hours). The trajectory of *degU* belongs to cluster 9 (Figure 3.10). The profile for *degS* corresponded to that in cluster 5 (Figure 3.9), the activity had two peaks, one of 330 TPM at the transition point and a maximum of 406.3 at 19 hours. In the case of *degQ* the transcriptional activity gradually increased from 149.5 TPM until a maximum of 822.4 TPM at the last sampling point (Figure 3.10 and Table 3.23).

For *B. licheniformis degU* did not peak at the transition point like in *B. pumilus*, rather the transcriptional activity showed a strong up shift at 4 hours but reached the maximal TPM of 1086.2 at 19 hours (cluster 20, Figure 3.14). *degS* followed the pattern described by cluster 7 (Figure 3.12) in which transcriptional activity increases to a maximum TPM of 192.8 at 7 hours and then decreased. Another difference with *B. pumilus* is the low abundance of *degQ* transcripts in regards to those of *degS* and *degU*. In *B. licheniformis*, *degQ* trajectory corresponds to cluster 11 (Figure 3.13), in which the transcriptional activity is downshifted for the first three sampling points with TPM values below 50, and then increments at 19 hours reaching 73.1 TPM (Table 3.24).

Transcriptional profiling points to DegU being more stimulated in *B. pumilus* than in *B. licheniformis* (Tables 3.23 and 3.24). In the case of *B. pumilus* there is a two-fold abundance of *degQ* transcripts in relation of *degS* at 7 and 19 hours, this could result in higher phosphorylation of DegU, which in turn might lead to more cells differentiating into the so-called "miner" type (producer of extracellular enzymes) [204]. In contrast, for *B. licheniformis degS* transcripts were more abundant at those time points, exceeding *degQ* by 7.1 and 1.9 ratios.

Regarding potential sRNAs targeting components of this regulatory system, three candidates were predicted for *B. pumilus* MS32 (Table 3.23). *degU* mRNA seem targeted by srna18 and its transcriptional activity presented a strong positive correlation, however it was not significant (0.873, pval=0.127). Two sRNAs were predicted to interact with *degS* transcripts, srna1 and srna17. The candidate sRNA17 corresponds to the Bpsr70 sRNA predicted for *B. altitudinis*, and it shows a positive but not significant correlation with *degS* of 0.722 (pval=0.278). No interacting sRNA partners were predicted for *degQ* mRNA.

For *B. licheniformis* there were seven sRNAs predicted to target *degU* mRNA (srna8, srna39, srna40, srna45, srna50, srna65, srna68), four for *degS* mRNA (srna19, srna32, srna65, srna70) and one for *degQ* (srna31) (Table 3.24). Most of the predicted sRNAs had a negative correlation with its target, except srna8, srna50 and srna31. The candidate srna8 had a correlation of 0.954 with *degU* transcriptional activity (pval=0.046), with both depicting maximal TPM at 19 hours. In the case of srna39 there was a strong negative correlation is -0.993 (pval=0.007). The candidate srna39 appears to be relevant only at the start of the fermentation, as the maximum TPM activity 114 was detected at 2.5 hours and then decreased to values below 5 TPM for the remaining sampling points.

Given the influence of the Deg system in the production of extracellular enzymes, it has caught the attention as an optimization target for industrially relevant strains. Hypersecretion mutants (with mutations known as *hy*), for example *degU32*, produce artificially high DegU-P levels in the cells, which result in higher protease secretion [33]. Upregulation of *degU-degS* has also been associated with higher protease yield in *B. altitudinis* [199]. Additionally, DegQ has been used as an enhancer of pullulanase production in *B. subtilis* WB800 [74]. Moreover, the system also impacts lipopetide production, knock out of *degU* had a positive effect on bacillomycin D and fengycin synthesis in *B. amyloliquefaciens* fmbj [307]. Further detailed examination of the transcriptional profiles of members of the DegU regulon could revealed the impact of the differences observed between *B. pumilus* MS32 and *B. licheniformis* MW3 and such insight might be used for strain optimization.

Spo0A

Another well characterized regulator in *B. subtilis* is Spo0A, the master regulator for sporulation initiation. Spo0A directly interacts with 143 genes according to Subtiwiki [250] but its modulating impact reaches around 500 genes [319]. Phosphorylation of Spo0A (Spo0A-P) is the result of quorum sensing signals and phosphorelay activation in which several histidine kinases and phosphatases interplay. The activated Spo0A-P binds to its target sequences which result in transcription activation or repression. SpooA represses the transition state regulator AbrB, moreover its activity impacts biofilm formation, cannibalism and competence development [319].

In *B. pumilus* *spo0A* transcriptional activity is represented by cluster 10 (Figure 3.10), which is down shifted at 2.5 and 4 hours, reaches a maximal TPM of 1033.9 at

7 hours and remains high until the end of fermentation (1066.9 TPM). For *B. licheniformis* the pattern of increasing transcription is similar (cluster 16, Figure 3.13), however the maximal activity is reached until 19 hours (1232.8) with a two-fold increase from the previous sampling point (632.5 TPM). Which points to differences in transcription and regulation, even in such a conserved system between these two closely related *Bacillus* species and how the development of spores might be timely different for them.

From the predicted candidate sRNAs in *B. pumilus*, six could potentially target *spo0A* transcripts (srna20, srna23, srna29, srna30, srna31 and srna40). A positive correlation of 0.835 (pval 0.165) was found between srna29 and *spo0A* transcriptional activities with both features reaching peak transcription at 7 hours (Table 3.23). In the case of *B. licheniformis*, three sRNA candidates (srna24, srna33 and srna68) are potential interaction partners of *spo0A* mRNA (Table 3.24). From those, srna68 had a negative correlation of -0.854 (pval 0.146) with the putative target, which seem to have a stronger impact during early time points of the fermentation as the sRNA was not detected anymore at 7 and 19 hours.

Given the impact of Spo0A in regulatory circuits and developmental changes mainly associated with stationary growth, it has been investigated for its effect in the production of industrially relevant products that take place during this phase. For example, it was described that deletion of *spo0A* resulted in loss of AprE synthesis capacity, while its overexpression lead to increased AprE transcription and activity in *B. licheniformis* 2709 [373, 374].

ComA

The master regulator ComA is a transcription factor that coordinates the regulation and development of the competence program in *B. subtilis*. The kinase ComP phosporylates ComA, phosphorylated ComA-P promotes the differentiation into surfacting-producing and competent cells [204].

For *B. pumilus* MS32 *comA* and *comP* followed the transcriptional trajectories depicted in clusters 11 and 9 respectively (Figure 3.10 and Table 3.23). While *comA* reached maximal transcription of 300.9 TPM at 19 hours with a peak of 220.6 TPM at the transition point, *comP* transcripts were more abundant at 4 hours (254.1 TPM).

Three sRNA candidates potentially target *comA* transcripts in *B. pumilus* MS32, srna0, srna3 and srna17 (Table 3.23). These three sRNAs respectively correspond to Bpsr193, Bpsr178 and Bpsr70 previously reported in *B. altitudinis* [353], although their physiological role remains to be elucidated. Both srna3 and srna17 showed peak activity at 19 hours, and a positive but not significant correlation with *comA*. The candidate srna5 seemed to target *comP* with a positive correlation and maximal TPM of 282.8 at 7 hours.

Interestingly, a comP homologue was not identified in *B. licheniformis* MW3 by the Proteinortho [188] analysis. The profile for *comA* corresponds to that in cluster 15 (Figure 3.13, Table 3.24), which reached maximal transcription activity at the transition point (83.4 TPM) and remained lower for the remaining time points. Two candidate sRNAs appear as possible interaction partners for *comA* mRNA, srna12 and srna29.

A significant increase in yield of iturin A was reported in *B. subtilis* ZK0 when *comA* and *sigA* were overexpressed, with the additional benefit of inhibiting biofilm formation [370]. If the regulatory roles of the candidate sRNAs are confirmed, they could contribute in similar optimization strategies for *B. pumilus* strains of industrial interest.

AbrA-Abba

The global transition state regulator AbrB interacts with genes determining different cell fates such as sporulation, competence, motility, degradative enzyme production, and biofilm formation. There are around 250 genes repressed by AbrB, which are negatively regulated during growth in favorable conditions [356, 321]. AbbA functions as an anti-repressor by mimicking the DNA phosphate backbone which prevents AbrB from binding to their DNA targets [321].

In *B. pumilus* MS32 AbrB follows the transcriptomic trajectory depicted in cluster 4 (Figure 3.9, Table 3.23), which describes an up-shift during exponential growth followed by down regulation after the transition point. For *B. licheniformis* AbrB is found in cluster 18 (Figure 3.14, Table 3.24) which follows a similar trajectory. In both cases the maximum transcriptional activity is reached at 4 hours, with 310 and 1211 TPM (triplicate mean) for *B. pumilus* and *B. licheniformis*, respectively.

Regarding Abba, it belongs to cluster 10 for both *Bacillus* (Figures 3.10 and 3.13). These clusters differ, for *B. pumilus* the trajectory increases gradually until the maximum transcriptional activity at 19. In the case of *B. licheniformis*, Abba is downshifted from the 2.5 time point to 4 hours and then it increases more sharply. More interestingly, the ratio between AbrB and Abba also differ for these two *Bacillus*, particularly at the first two sampling points. While *B. pumilus* has a AbrB/AbbA ratio of 0.84 and 0.98, *B. licheniformis* presents a ratio of 9.07 and 16.34, at 2.5 and 4 hours, respectively. This could point to differences in the regulatory pathways of this master regulator.

Actually, such an important regulator is under sophisticated controls systems that fine-tune its activity. AbrB is also repressed by Spo0A [356], and more recently it was found that three kinases can phosphorylate AbrB to prevent its DNA binding function [178]. Perhaps sRNAs also play a role in this regulatory system. In *B. pumilus* none of the putative sRNA had a predicted interaction either with AbrB or AbbA. However, in *B. licheniformis* two sRNA candidates (srna34 and srna43) could potentially target AbrB and other two other (srna13 and srna25) AbbA. The predicted srna25 has a correlation of 0.976 (pval 0.024) with *abbA* transcriptional activity, and the calculated energies for this interaction are -15.99, -15.46 and -134.694 by RNAplex, RNAup, IntaRNA as implemented within Annogesic [363], respectively.

AbrB has been used as a production optimization target. In the genome-reduced *B. subtilis* MGB874 the constitutive expression of AbrB released the cell from the burden of expressing the AbrB-regulated genes that lead to differentiated cell types [356]. By maintaining the cell population in a undifferentiated state the bioproductivity was increased, as well as the metabolic flux of glycolysis and other pathways associated to a growth defect of the strain. Production of poly-γ-glutamic acid improved, as well as the extension of the productive phase of cellulase compared to the 168 strain [356].

ScoC

The transition state regulator ScoC has a high impact on *Bacillus* physiology. It is involved in motility, competence, membrane transport, oxidative stress response, sporulation and motility [43], Subtiwiki [250] indicates 39 genes regulated by ScoC. Moreover, it has been reported to negatively regulate the transcription of the protease encoding *aprE* [20, 43].

Transcriptional activity of *scoC* follows different patterns in *B. pumilus* and *B. licheniformis* (Tables 3.23 and 3.24). For *B. pumilus scoC* transcription gradually increases from 100.3 TPM until a maximum of 317.9 at 7 hours followed by a three fold downshift at 19 hours (cluster 7, Figure 3.9). *scoC* was reported to have peak activity during exponential growth and then declined in *B. pumilus* BA06 [129], which is different from the observed data. In *B. licheniformis scoC* starts with high activity at 2.5h, a sharp decrease at the transition point, reaching a maximal transcription at 7 hours with 982.8 TPM and then it is down shifted at 19 hours, although not as strongly as in MS32 (cluster 4, Figure 3.12). This points to interspecies differences in regulation and adaptation to the fermentation environment, particularly at earlier time points.

Higher levels of *ScoC* in the middle of the productive phase could have detrimental effects on the yield of industrial bioprocesses. However, the intricate regulatory networks with often redundant systems in place, complicate the scenario. In *B. subtilis*, *ScoC* is activated by AbrB and SenS (no SenS homologue was found in *B. pumilus* MS32 or *B. licheniformis* MW3 by the ProteinOrtho [188] analysis) and repressed by CodY and SalA (both present in the studied *Bacillus* strains). A study on regulators of extracellular proteases in *B. subtilis* found that increased expression of *aprE* was only achievable in a triple mutant with inactive CodY, ScoC and AbrB [20]. In contrast, higher extracellular protease activity was found for a *B. pumilus* BA06 mutant with disrupted *scoC* [128]. None of the predicted sRNAs seem to interact with ScoC transcripts in both of the *Bacillus* species studied.

CodY

CodY is another major transcriptional regulator, sensing levels of branched chain amino acids (BCAAs) and GTP molecules. When their concentration decreases in nutrition limiting conditions, derepression of CodY targets occur (for example rapA, phrA, kinB, phrF, rapF [36]). This derepression activates systems leading to bacterial adaptations to regulate nutrients and energy metabolism [319]. CodY also interacts with TnrA (a regulator of nitrogen metabolism) and CcpA (regulator of carbon metabolism) and has an indirect effect on synthesis of extracellular proteases by interacting with ScoC [20, 319].

In *B. pumilus* MS32 CodY transcripts were highly abundant at 2.4 and 4 hours, then they decreased at 7h, followed by a slight up shift at 19 hours (cluster 4, Figure

3.9, Table 3.23). In contrast, for *B. licheniformis* CodY transcriptional activity peaked at the first time point with a stronger down shift occurring at 4 hours, and a similar up shift towards the final sampling point *cluster 14, Figure 3.13, Table 3.24*. The regulatory impact of opposite CodY transcriptional activities between these *Bacilli* at the transition point is to be determined, and it requires careful evaluation given the multiple interacting regulators associated with its effect. The earlier downregulation of CodY in *B. licheniformis* could point to a more rapid exhaustion of nutrients and earlier activation of adaptive responses, such as activation of alternative metabolic pathways. An additional layer of regulation mediated by sRNAs could not be identified for the studied *Bacilli* species.

CsrA

The carbon storage regulator A CsrA has a post-transcriptional regulatory function impacting flagellar morphogenesis. In *B. subtilis* CsrA binds to the flagellin *hag* transcript and occludes the Shine-Dalgarno sequence, impeding *hag* translation [231]. Additionally, a RNA chaperone role has been described for CsrA, in which it facilitates the interaction of the sRNA SR1 with its *ahrC* target mRNA, a transcriptional regulator involved in arginine metabolism [232, 323].

The transcriptional profile of *csrA* differs between *B. pumilus* MS32 (cluster 9, Figure 3.10, Table 3.23) and *B. licheniformis* MW3 (cluster 17, Figure 3.14. Table 3.24) during the fermentation. For MS32 transcriptional activity is down-shifted at 2.5 hours and then increases reaching a TPM peak of 445 at 7 hours maintaining the up-shift until 19 hours. On the other hand, for *B. licheniformis* it starts with high abundance from the start of fermentation and reaches its maximum of 379 TPM at 4 hours followed by a strong decrease instead. There was no sRNA candidate predicted by Annogesic [363] with homology to SR1. However, according to the detection based on the highly sensitive covariance models of RFAM both *B. pumilus* MS32and *B. licheniformis* MW3 genomes could encode a RNA candidate with homology to SR1 (Table 3.14), which could be further examined in regards with their regulatory role.

Synthesis and assembly of flagellar components is a energy-expensive process, it has been reported as particularly high for *B. subtilis* which can synthesize around 20 flagella per cell [231]. If the transcriptomic profile of *csrA* in *B. pumilus* MS32 related to less energy invested in flagellar production, this could allow the cells to direct resources to growth and probably to synthesis of proteins of industrial interest. In fact, when the *fla* operon in *B. subtilis* was knocked out by CRISPR-dCas9 methods, the production of amylase was significantly increased [89].

SwrA-SwrB

SwrA is known as the master activator of flagellar synthesis, it impacts the *fla-che* operon by interacting with DegU-P. Within the operon *sigD* and *swrB* are also found. More recently the regulatory functions of SwrA were extended beyond motility, as it affects other DegU targets with important consequences for competence and protease production [84]. SwrB is required for SigD activity and has a key function in the activation of the flagellar type III secretion export apparatus, mutants without

SwrB were found defective in swarming [256].

SwrA follows contrasting transcriptional trajectories in *B. pumilus* and *B. licheniformis* (cluster 17 in Figure 3.11, Table 3.23 and Figure 3.14, Table 3.24 respectively). For *B. pumilus* SwrA starts down-shifted at 2.5 hours, followed by an eight fold upshift at 4 hours and a peak activity of 1617 TPM at 7h, and remains abundant until the last time point. In contrast, for *B. licheniformis* the maximal TPM of 271.7 was detected at 2.5 hours and then the transcriptional activity was gradually decreased to 22 TPM at 19 hours.

The transcriptional profile of *swrB* was also different. For *B. pumilus* the transcriptional activity started high, peaked at the transition point (242.6 TPM) and then was decreased (cluster 18, Figure 3.11). While in *B. licheniformis* the peak occurred at 2.5 hours (334.08 TPM) followed by a strong three-fold down shift at 4 hours and continued to decrease until the end of the fermentation (cluster 6, Figure 3.12).

The evidence points to these *Bacilli* eliciting different responses to the fermentation environment which triggered the regulatory and developmental changes associated to motility in different time points. A recent report expanded the interaction network of SwrA, and found that it also modules DegSU, which lead to a reduced transcription of *aprE* as well as decreased secretion of cellulases and xylanases in *B. subtilis* [84]. Here, *swrA* transcripts were highly abundant at 7 hours for *B. pumilus*, if the negative effect observed for *B. subtilis* also applies for MS32, this could point to a potential optimization target.

An additional level of regulation could take place regarding *swrA* and *swrB* activities. For *B. pumilus* 2 candidate sRNAs potentially interact with *swrA* transcripts (srna30 and srna31, the putative SsrA) and 3 with *swrB* mRNA (srna12, srna 39 and srna42) (Table 3.23). The candidate srna39 was found to have a correlation of -0.905 (p-val 0.095) with *swrB*, the interacting energies predicted by RNAplex, RNAup, IntaRNA were -17.36, -7.99 and -919.545, respectively. In *B. licheniformis* no candidate sRNA was predicted to interact with *swrB*, while srna28 and srna71 could potentially target *swrA* (Table 3.24).

SinR/I SlrA/R

In *B. subtilis* SinR is the primary transcriptional regulator coordinating biofilm formation. During growth favorable conditions, it exerts negative regulation over the *epsA-O* and *tapA-sipW-tasA* operons, which are essential for matrix production [225, 165]. Upon environmental changes to unfavorable conditions, quorum sensing signaling triggers the production of SinI. SinI forms an heterodimer complex with SinR which prevents its DNA binding function and leads to derepression of SinR targets [165]. SinR also represses *degU* [242]. More recently, SlrR and SlrA were discovered as paralogues of SinR and SinI, respectively, SlrA is a small peptide antagonist to SinR. SinR represses the expression of *slrR* [177].

In *B. pumilus sinR* and *sinI* transcriptional activities follow cluster 17 (Figure 3.11). However the abundance of *sinR* transcripts is 8 times higher than *sinI* at 2.5 hours and remains three fold increased during the remaining time points. Both genes reached maximal TPM of 450 and 146, respectively, at 19 hours. Transcription of *slrA* was rather down shifted during the first three sampling points and only peaked with 140.5 TPM at 19 hours. Interestingly, *slrR* transcripts were very scarce during the fermentation, the maximal detected TPM was 11 at 7 hours, TPM values were below 10 for the rest time points (Table 3.23).

From the candidate sRNAs, srna36, srna10 and srna19 were predicted to interact with *sinR*, *SinI* and *SlrA*, respectively. No interacting partner was found for *slrR*. srna10 corresponds to the previously proposed Bpsr139 and a positive correlation (0.544, pval 0.456) with SinI is suggested, although it has low significance, the interaction is supported by independent RNAplex, RNAup and IntaRNA analysis within Annogesic [363], with an interaction energy of -221.331 predicted by IntaRNA. Additional experiments are required to confirm the potential targets of these sRNAs and their putative role within this regulatory circuit.

For *B. licheniformis*, *sinR* also reached peak activity at 19 hours with 311.8 TPM and follows the trajectory described by cluster 16 (Figure 3.13) in which transcription is downshifted until 7h. *sinI* was found in cluster 3 (Figure 3.12) and unlike *B. pumilus* the transcriptional activity was rather low with a maximal TPM of 44.6 at 19 hours and activity below 15 TPM in the remaining time points. Similar to *B. pumilus sinR* transcripts were found in excess compared to *sinI*, however SinR seems even more abundant for *B. licheniformis*, with a SinR/SinI ratio of 22.3 at 2.5 hours and between 7 and 15.2 at the other time points. *slrR* appeared also low and was not detected at 19 hours, *slrA* reached a peak at 4 hours with 96.2 TPM (Table 3.24).

More candidate sRNAs could potentially interact with these regulators in *B. licheniformis*. srna51 and srna61 (both matching RFAM:RF00168) are predicted partners of *sinR*, while *sinI* was potentially targeted by: srna40, srna47, srna51 and srna61. For *slrA* srna43 and srna64 are proposed as interacting sRNAs, as well as srna62 for *slrR*. The putative srna40, matched the RFAM:RF00011 entry, which describes the ribozyme Ribonuclease P (RNase P). In *B. subtilis* the only reported targets of RNase P are precursor tRNAs, for which RNase P has a role in maturation. However, due to RNase P ability to associate with 30S ribosomal subunits, it has been suggested that it could cleave regulatory regions in mRNAs in order to induce translation [25]. If the predicted interaction is confirmed experimentally this report could expand the target range of RNase P.

Further analysis opportunities. The generated genomic and transcriptomic data could be additionally interrogated and combined to generate significant biological insights. For example, a pipeline integrating identification, inter-species comparison, and analysis of candidate sRNAs and their potential targets within an interactive visualization interface for easier interpretation was devised as a master thesis project. The pipeline was developed and implemented by Anton Robert Georg Farr using this project's data [87]. Figure 4.1 presents an overview of the resulting pipeline, which is currently under further development.

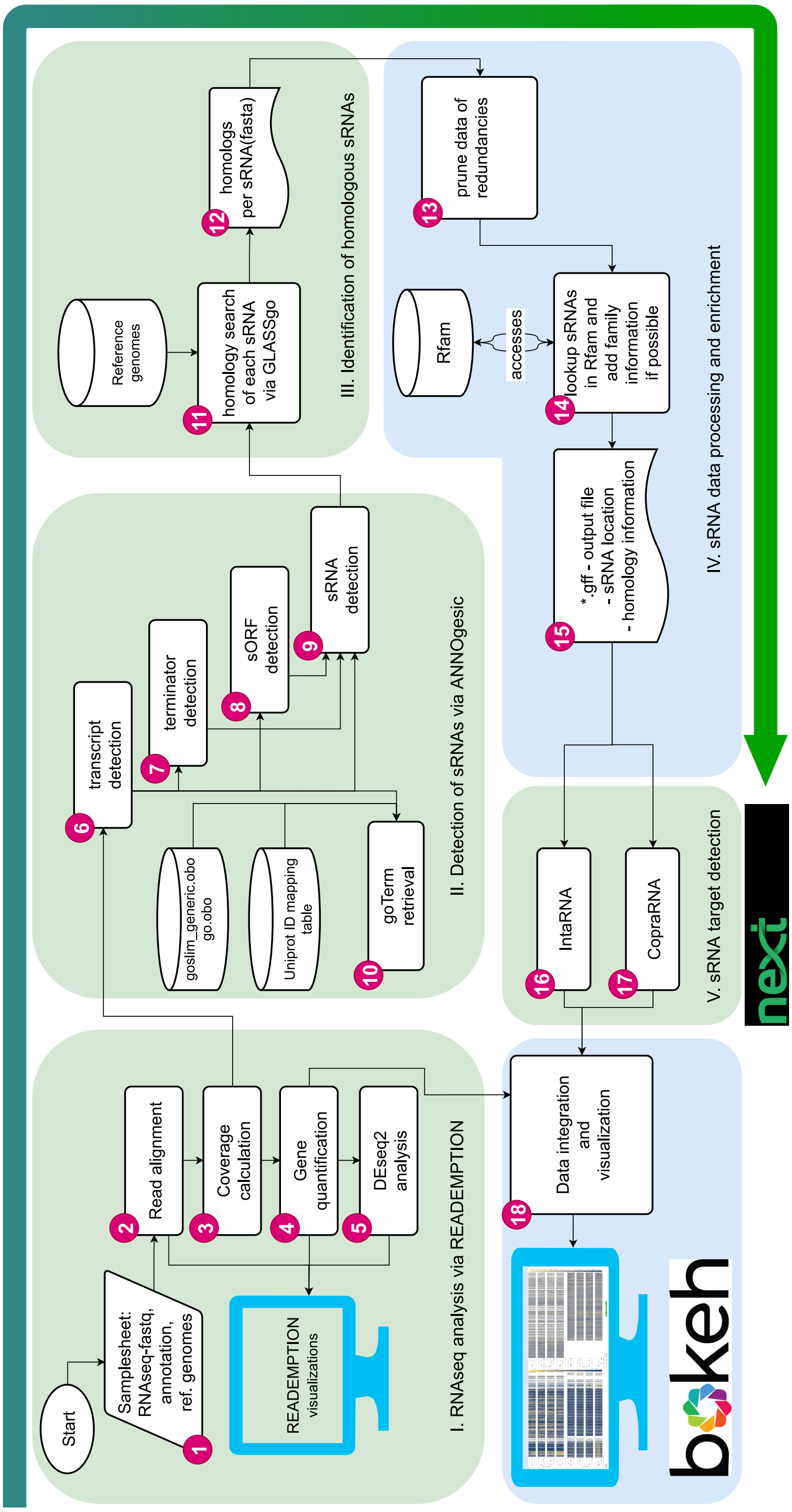


FIGURE 4.1: Overview of a bioinformatic pipeline for identification and comparison of transcriptionally active sRNAs and their target mRNAs [87]. The pipeline processes RNA-seq data with Annogesic [363] and READemption [104], identifies sRNAs, mRNA targets, scans closely related genomes for homologous sequences, performs functional enrichment, and integrates the results in an interactive report for exploratory analysis [87]. Figure provided by Anton Farr.

4.5 Conclusions and Outlook

The metabolically versatile genus *Bacillus* offers promising hosts for biotechnology industry applications. Many of their natural adaptations can be exploited for increased productivity. At the same time, some of their innate features (for example, sporulation and Restriction-Modification systems) can also hinder the transition from wild strains to productive industrial workhorses. By having good genetic accessibility, a strain of interest can be studied and modified for indsutrial purposes. Several methods for bacterial transformation have been developed, but adapting the protocols for recalcitrant strains is still challenging. This was evidenced by the different failed attempts to transform *B. pumilus* DSM27.

This work presented a comparative genomic analysis of *B. pumilus* strains, together with two *B. subtilis* and *B. licheniformis* representatives. The first portion of analysis showed that the novel isolate MS32 corresponds to the *B. pumilus* species, and it is close to the type strain DSM27. Therefore it was a suitable candidate for fermentation experiments and further transcriptomic studies.

A comprehensive characterization of *Bacillus* genomes in terms of prophages, Insertion Sequence elements, CRISPR-Cas systems, antibiotic resistance determinants, Restriction Modification systems, proteases, secretion machinery, and clusters for biosynthesis of secondary metabolites was presented. This provided an overview relevant features of interest during the development and optimization of productive hosts. For instance, the identified RM systems offered an explanation for DSM27 limited genetic accessibility. Moreover, unexplored potential in *B. pumilus* regarding bioactive compounds was highlighted. For example the putative novel class III lanthipeptide and the NRP-Polyketide type of product which seemed highly active, but remains to be characterized. Similarly, there is a repertoire of enzymes with promising applications, such as the cyanide dihydratase which could be used for detoxification purposes.

Bacillus species have the capacity to exploit a diverse range substrates. Their robust and versatile metabolism allows the processing of cheap carbon and nitrogen sources such as agro-waste materials, which is in line with sustainable production goals. The comparison of metabolic maps indicated that the main pathways observed in *B. licheniformis* and *B. subtilis* are also present in *B. pumilus*, which points to similar metabolic capacities. Nevertheless, *B. pumilus* transport systems, particularly related to amino acid transport and metabolism were more abundant. Furthermore, regarding iron acquisition mechanisms, *B. pumilus* appears to be able to exploit additional iron sources, and genes for those transporters were highly transcribed.

Comparative genomic analysis could be expanded to add more genomes of *B. pumilus* and other *Bacillus* species. Additional analysis could include determination of: core and pangenomes, genomic islands, carbohydrate active enzymes, and determination of virulence factors against specialized databases.

An optimized protocol for RNA isolation from industrially relevant *Bacillus* samples was developed. The modifications generated RNA of high quality, purity and integrity which translated into high quality RNA-seq libraries. Moreover, the protocol could be transferred to other bacteria of biotechnological interest from which standard isolation methods fail to deliver RNA of optimal quality.

By studying *Bacillus* organisms in fermentation conditions, not only we gain insight of the biological traits that allows them evolutionary success in a wide range of environments, but we also generate knowledge that aids in bioprocess engineering strategies. A collection of clusters describing the transcriptional profiles of genomic features of *B. pumilus* and *B. licheniformis* across the fermentation time points was generated. This could be integrated with the DEG analysis output to identify significant changes in those transcriptional profiles. This would point to which elements are most (significantly) relevant at different growth phases and potentially form part of shared regulatory networks. Additional analysis could include identification of regulatory recognition motifs within clusters to further elucidate these networks.

The comparative transcriptomic analysis allowed to investigate *B. pumilus* and *B. licheniformis* responses in conditions similar to industrial fermentation processes. A catalog of transcriptionally active sRNAs (including known and putative novel ones) was identified for *B. pumilus* and *B. licheniformis*. They were characterized and their transcriptional profiles were determined. Moreover, candidate mRNA targets were proposed. These results could be used for global coexpression analysis and inference of regulatory networks showing the full range of targets of a given candidate sRNA and its potential regulatory effect.

Proteases and the secretory machinery of *B. pumilus* and *B. licheniformis* were characterized by determining and comparing their temporal expression profiles and putative interacting sRNAs. Knowledge on such elements is beneficial to solve bottlenecks in productive processes. This work also elucidated potential interactions with regulatory effects at different phases of the fermentation, particularly regarding known global regulators coordinating major cellular processes, metabolism and features of industrial interest like production of degradative enzymes. For example, it was discovered that DegU was the most actively transcribed regulator (from the studied set) in *B. pumilus*, and sRNAs interacting with *degU* transcripts were suggested. The proposed interactions need to be experimentally validated, by Northern Blot, for example. Once characterized, novel sRNAs could be integrated in the creation of covariance models and submitted to resources like RFAM, which in turn would aid in the identification of more sRNAs in other bacterial genomes.

By generation of cDNA libraries from the isolated RNA of *B. subtilis* (which were collected from parallel fermentation runs and processed in the same way as the *B. pumilus* and *B. licheniformis* samples), the comparative transcriptomic analysis could also be expanded. Another interesting line of study could be the determination of methylation profiles for these *Bacillus* species. This could add another relevant layer of regulation with impact for biotechnological applications.

In conclusion, investigation on *Bacillus* organisms as microbial cell factories is required for a transition towards a circular bioeconomy. This study characterized *B. pumilus* from comparative genomics and transcriptomics perspectives. The data integration from the selected approaches demonstrated to be useful to generate relevant insights. Therefore, does *B. pumilus* represent a bacteria with potential as a high performance expression platform? *B. pumilus* shares many relevant and desirable features with species offering established cell factories, such a secretion machinery, versatile metabolism, fast growth, and robustness in fermentation conditions. Additionally, it appears superior in amino acid and iron transport systems. It also offers novel compounds for yet uncharacterized applications. Transcriptomic data contributed to identify major regulatory nodes impacting cell performance, particularly sRNA regulators. *B. pumilus* offers a good candidate for biotechnological applications. This study not only contributed to the understanding of *Bacillus* biology, but also promises to support the development and optimization of *B. pumilus* as a high-performance expression platform.

Chapter 5

Science communication, "secreting" discoveries outside academia

Chapter 1 contextualized investigations on *Bacillus* industrial production platforms as part of a bioeconomy-oriented society framework. Therefore, there is an additional aspect to explore regarding the research presented so far: science communication. This is an element often overlooked in research, however it has profound impacts for society in general. This chapter presents a brief commentary on science communication, its diverse functions, challenges, benefits, and overall relevance. As a small application case, perception indicators about *Bacillus* and biotechnology were collected from internet and a graphic communication product is proposed as a way to address those impressions.

5.1 Sharing the passion

Despite some perceptions of science as impersonal, objective and evidence-oriented, scientific work involves passion and curiosity. The amount of hours, weekends, years, and resources devoted to the investigation of an organism, a question, or method, reflects it. Passion fuels daily efforts and creativity, while promoting resilience in the face of adversity [147]. For a scientist, sharing that passion can go beyond publications in specialized journals, conference presentations, and lecture halls. There is a whole world of possibilities for science communication with broader and diverse audiences.

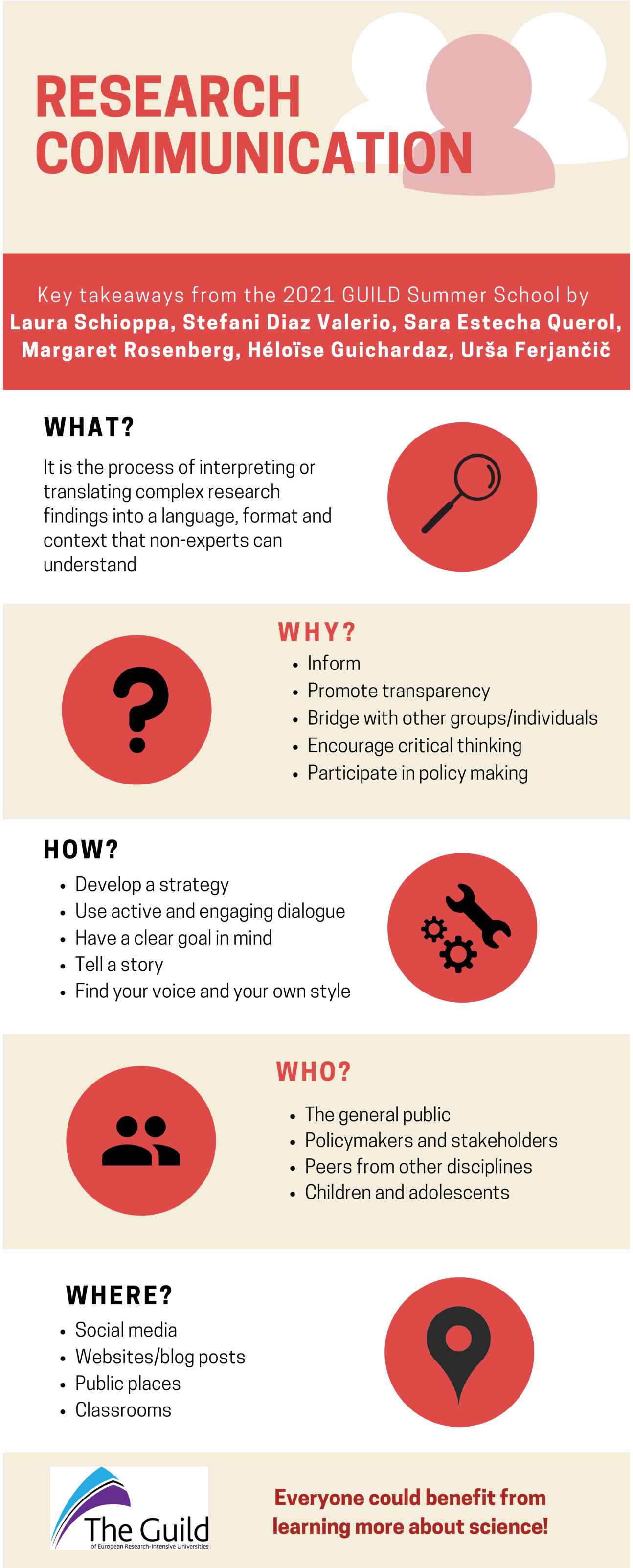
There are discussions around the definition of science communication (SciCom) and how it relates to other concepts (often used interchangeably), such as: public understanding of science, scientific culture, public engagement, scientific literacy, research communication, public communication of science and technology (PCST), public awareness of science and science journalism [40, 166, 246]. For the purposes of this chapter, which is not meant as an exhaustive review on the topic or a technical discussion into communication paradigms, science communication can be understood as an umbrella term. This therm, depending on the balance of factors such as: the target audience, goal, media, message, and communication strategy, could be narrowed to more specific terminologies.

Figure 5.1 presents general aspects of what can be understood as science communication based on the so-called journalistic questions. The figure was produced at the Guild Summer School 2021, hosted by the University of Glasgow. The school was promoted by the Guild of European Research-Intensive Universities and was aimed at PhD candidates from multiple disciplines and from across Europe. The graphic product was created by non-experts in communication, rather by future researchers from different areas, and therefore lacks some of the nuances that would allow a more thorough description. A comprehensive definition is presented in the following box.

"Science communication (SciCom) may be defined as the use of appropriate skills, media, activities and dialogue to produce one or more of the following personal responses to science (the vowel analogy)

- Awareness, including familiarity with new aspects of science
- Enjoyment or other affective responses, e.g. appreciating science as entertainment or art
- Interest, as evidenced by voluntary involvement with science or its communication
- Opinions, the forming, reforming, or confirming of science-related attitudes
- Understanding of science, its content, processes, and social factors

Science communication may involve science practitioners, mediators, and other members of the general public, either peer-to-peer or between groups" [40].



RESEARCH COMMUNICATION

Key takeaways from the 2021 GUILD Summer School by **Laura Schioppa, Stefani Diaz Valerio, Sara Estecha Querol, Margaret Rosenberg, Héloïse Guichardaz, Urša Ferjančič**

WHAT?

It is the process of interpreting or translating complex research findings into a language, format and context that non-experts can understand

WHY?

- Inform
- Promote transparency
- Bridge with other groups/individuals
- Encourage critical thinking
- Participate in policy making

HOW?

- Develop a strategy
- Use active and engaging dialogue
- Have a clear goal in mind
- Tell a story
- Find your voice and your own style

WHO?

- The general public
- Policymakers and stakeholders
- Peers from other disciplines
- Children and adolescents

WHERE?

- Social media
- Websites/blog posts
- Public places
- Classrooms

Everyone could benefit from learning more about science!

FIGURE 5.1: Infographic presenting an overview of research communication as produced within the 2021 Summer School of the Guild of European-Research-Intensive Universities.

As shown in figure 5.1, there are many reasons to engage in SciCom activities. For example, given the amount of public funding dedicated to research, accountability is one of the functions of SciCom. As scientists, it is part of our responsibilities to share with the public the results (and impacts for society) of the investigations that are funded, otherwise from the public's perspective, the investment lacks justification, interest is lost, and funding decreases.

Moreover, science communication is necessary for key actors to make informed decisions. From the general public in day to day interactions, to policy makers, stakeholders, investors, companies, and institutions, these are all agents who contribute in the creation of a society agenda of priorities. The relevance of taking scientific knowledge out of the ivory tower is increasingly recognized, notably from academia itself. For instance, a recent publication [316] made an urgent call for microbiology literacy in society. Given the pervasive and profound impact of microbes in our lives (from the human microbiome and industrial activities, to the entire biosphere scale), the authors proposed to include microbiology in the public education curricula and envision a Microbiology literacy framework [316].

To inform and inspire, that is another reason to participate in SciCom. As evidenced by several surveys and studies, there is a great appetite and increasing interest in science by society in general. For instance, a 2015 report by the Welcome Trust found 63% of the people surveyed interested in hearing about scientific discoveries directly from researchers [56]. Moreover, as scientists, there is a responsibility, an ethical compromise, to engage and nurture the next generation of scientists, this can take place even before they arrive to university campuses. By participating in SciCom activities, we can inspire young individuals to pursue STEAM paths. For example young girls, and consequently contribute to increased female representation in scientific areas.

The recent COVID-19 pandemic is a great example evidencing the need and the importance to inform about science. Suddenly, there was a high demand to know about viruses and the immune system by general public and audiences who needed the information but lacked specialized training. Information about everyday scientific methods, such as PCR and sequencing, was necessary for people outside academia in order to make informed decisions that had deep repercussions on public health policies. Conversely, the pandemic also highlighted the risks and consequences of misinformation, which lead to conspiracy theories and the spread of non-scientific views. Institutions, like the World Health Organization, made a call for scientist to serve as communicators and help combat misinformation trends [5, 24]. Were scientists ready to answer the call? Will we be prepared enough for the next time?

Undoubtedly, research is labor-intensive, demanding, and time-consuming. Scientists joggle between teaching, supervision, operative maintenance of research facilities, writing, presentations, grant applications and research itself. Barriers to SciCom by researchers include limitations regarding: time, funding, opportunity, and training, as well as lack of value recognition for SciCom engagement [56]. Therefore, formal incentives and acknowledgments are required if researchers are expected to add SciCom to their activities. Fortunately, this has been recognized, and nowadays research funding agencies are shifting to encourage SciCom elements in grant applications. For example, The European Union's Seventh Framework Program for

Research and Technological development promotes participants to "communicate and engage with actors beyond the research community" [246].

> “We are, as a species, addicted to story. Even when the body goes to sleep, the mind stays up all night, telling itself stories.”
> — Jonathan Gottschall, The storytelling animal: How stories make us human [120].

Even though the SciCom task might appear as challenging for scientists, there are intersections between SciCom and specialized scientific communication. After all, communication skills are also required within research, where scientist also assume storytelling roles. This ability is exercised when choosing the story frame of an article or a conference presentation, and Why? Because stories are engaging and a natural aspect of human interactions. Stories facilitate understanding and help to consolidate knowledge. Further elements of this intersection between SciCom and technical communication include: clarity, defined messages, coherence, and awareness of the target audience.

Moreover, there are direct benefits for researchers who engage in SciCom activities. Those include [56]:

- Development and improvement of communication skills.
- Appreciation of the research value from a non-academic perspective (which can impact study designs and aims).
- Satisfaction and fulfillment by interaction and exchange with the public.
- Expanded research impact (and visibility) by reaching broader audiences. This is of particular interest in a context moving from the "publish or perish" view to a "be visible or vanish" perspective. [22]
- Further exchange with specialists from different areas, which starts conversations with the potential to develop into exciting multidisciplinary projects.

SciCom can take many forms: science café, pint of science, science slams, science festivals, exhibitions, school visits, blogs, newsletters, videos, podcasts, infographics, magazine articles, comics, art, music, social media posts, games etc.

5.2 Case of study: *Bacillus* and biotechnology

As discussed in Chapter 1 the transition towards a circular bioeconomy is of great relevance to face current day challenges. Biotechnology strategies, like production, optimization, and implementation of enzymes produced by microorganisms play a key role within this transition. Therefore the study of microbial cell factories, such as *Bacillus* species, is of special interest. Nevertheless for a successful social transformation to a more sustainable model, this knowledge must be spread out of academia and reach policy makers, stakeholders, consumers, and society in general [170], therefore, science communication strategies are needed.

There are SciCom strategies already promoting *Bacillus* species. They are mostly focused on *B. subtilis*. For example, the microbe of the 2023 year by VAAM (Association for General and Applied Microbiology) is *B. subtilis*. The "Microbe of the Year" is an initiative to highlight the role that microorganisms play in ecology, health, nutrition, and economy. Similarly, a recent publication introduced *B. subtilis natto* applications for plant protection and in the production of natto (fermented soybean) [186]. The twist, the article was addressed to children instead of researchers. It was published in the journal Frontiers for Young Minds, which aims to make scientific discoveries available to younger audiences. They do so by encouraging scientists to write shorter and easy-to-comprehend versions of their usual manuscripts, these articles are freely available to educators around the world.

5.2.1 Perceptions around *Bacillus*

Creation of a novel environmentally friendly product or technology is not enough if the public, policy makers, and specially potentially consumers are not taken into account [170]. Communication in general is a two-way road, SciCom is not only about researchers telling broader audiences about their investigations and their impact, but it is also an opportunity for dialogue. There has been a shift from the "deficit model", a linear perspective that considers the public's knowledge as inadequate and their role as passive, towards a "contextual or participative model", in which there symmetry in the flow between science and its publics [40].

There are online tools to collect and summarized data about the most related search terms and questions associated with query words. This is helpful to gain understanding of public's perceptions, gaps, and attitudes around a given topic. By doing so, communication strategies can take place and be tailored to the target audience and its needs. Of course there are more focused tools aimed to specific groups of interest, for example targeted surveys and focus groups. Regarding general online inquiries, AlsoAsked `https://alsoasked.com/` gathers the information from the "People also ask" section of Google search results. And AnswerThePublic `https://answerthepublic.com/` collects data from Google's auto-suggest field. As a brief example, the prompts "Bacillus" and "Biotechnology" were given to AlsoAsked and "Bacillus" and "pumilus" were passed to AnswerThePublic. The Figures 5.2 and 5.3 present the result of those inquiries.

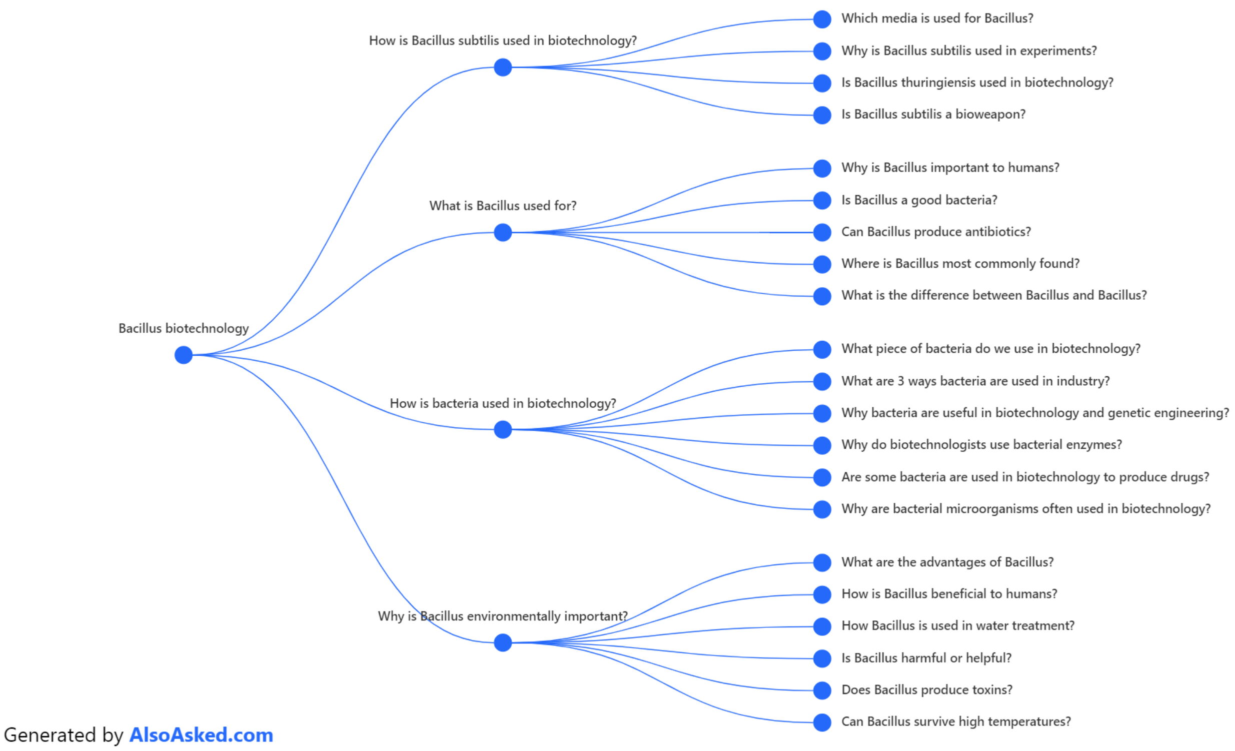

Bacillus biotechnology
How is Bacillus subtilis used in biotechnology?
Which media is used for Bacillus?
Why is Bacillus subtilis used in experiments?
Is Bacillus thuringiensis used in biotechnology?
Is Bacillus subtilis a bioweapon?
What is Bacillus used for?
Why is Bacillus important to humans?
Is Bacillus a good bacteria?
Can Bacillus produce antibiotics?
Where is Bacillus most commonly found?
What is the difference between Bacillus and Bacillus?
How is bacteria used in biotechnology?
What piece of bacteria do we use in biotechnology?
What are 3 ways bacteria are used in industry?
Why bacteria are useful in biotechnology and genetic engineering?
Why do biotechnologists use bacterial enzymes?
Are some bacteria are used in biotechnology to produce drugs?
Why are bacterial microorganisms often used in biotechnology?
Why is Bacillus environmentally important?
What are the advantages of Bacillus?
How is Bacillus beneficial to humans?
How Bacillus is used in water treatment?
Is Bacillus harmful or helpful?
Does Bacillus produce toxins?
Can Bacillus survive high temperatures?
Generated by AlsoAsked.com

FIGURE 5.2: Diagram summarizing the most common questions associated to the terms "*Bacillus*" and "Biotechnology" as generated by AlsoAsked https://alsoasked.com/

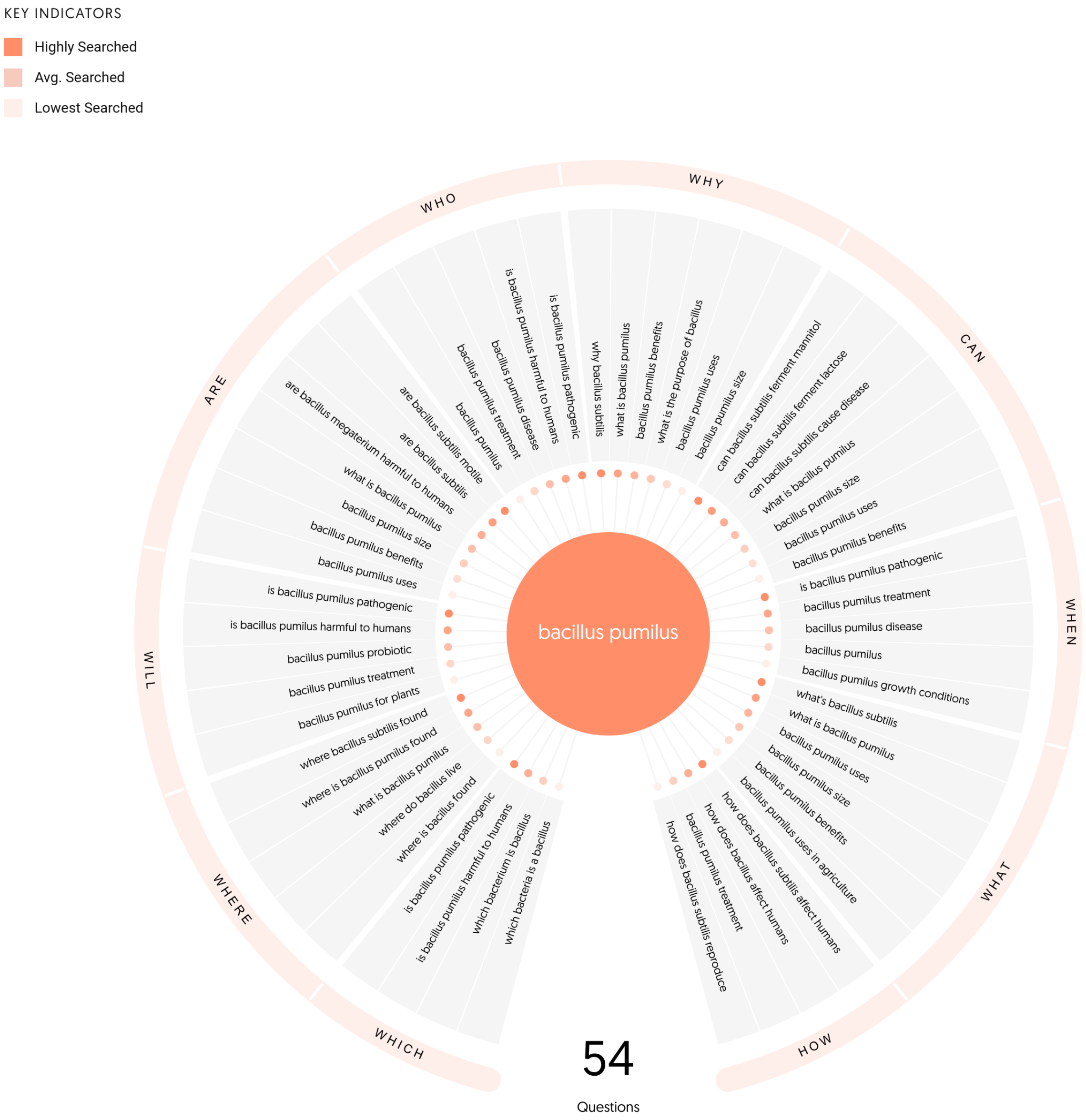


FIGURE 5.3: Summary of the most related questions around "*Bacillus*" and "*pumilus*" as collected by AnswerThePublic `https://answerthepublic.com/`

From the Figures 5.2 and 5.3 it is shown that some concerns around *B. pumilus*, and *Bacillus* in industry in general, are regarding biosafety, particularly pathogenic potential. For example in both figures variants of "Is *Bacillus* (*pumilus*) pathogenic?" and "Is *Bacillus* harmful or helpful?" are often found.

Noticeably, the figures also evidence interest ("How does *Bacillus* affect humans?" and "Why is Bacillus important to humans?"), and maybe a gap to very fundamental questions, such as "Where is *Bacillus* commonly found?". Interestingly, by searching the term "*Bacillus*" particular species were frequently found, for example: *B. subtilis*, *B. anthracis*, *B. megaterium* and *B. thuringiensis*. Perhaps those species already form part of public's perceptions and could be used as anchors to further introduce other organisms, like *B. pumilus*, for instance by reiterating what makes *B. pumilus* closer to safe-to-use organisms (*B. subtilis*) and what differentiates it from virulent ones (*B. anthracis*).

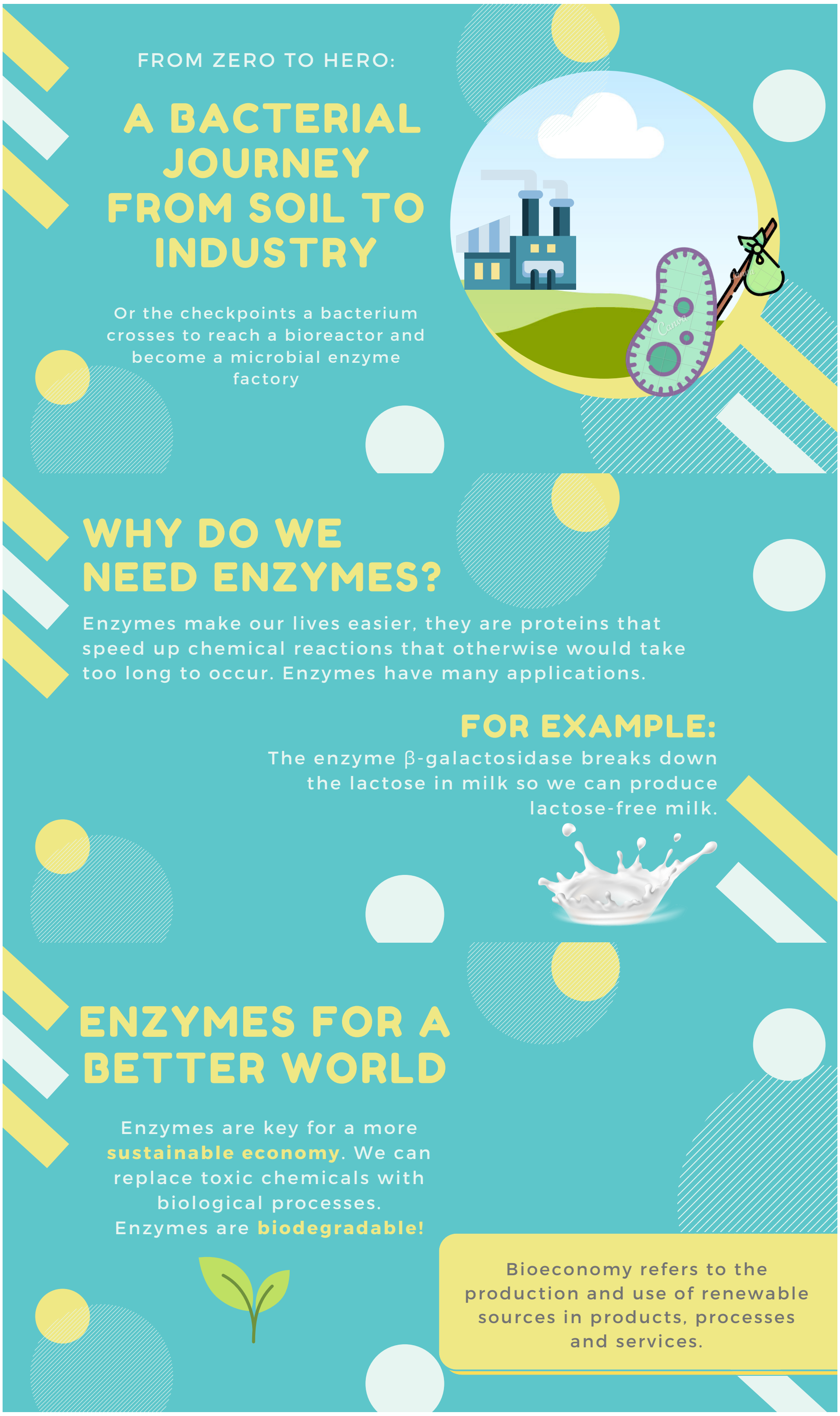
5.2.2 "From Zero to Hero" an infographic series about *Bacillus*

"Just like Bacillus species secrete enzymes and incorporate the processed products to support cell growth, science communication brings knowledge outside academic environments and returns relevant feedback to promote scientific advancements"

This work could be translated to broader audiences and used to address some of the concerns found in the previous section. The "From Zero to Hero" series could be used to present different stories around *Bacillus*. The first one presents "A Bacterial Journey from Soil to Industry". It starts by exposing the "Why". Why do we need enzymes?, Why are they relevant? It states a purpose within a bigger bioeconomy context, and uses common day examples (which are relatable for the lay audience). The examples of detergents, bread, and paper allow to make a direct more personal link than just data and numbers. After the why is presented and relatable examples are given, then *Bacillus* organisms are introduced. One of the common questions is answered: Where are *Bacillus* found? and are they safe to use for industrial applications? Then, some of the features that make *Bacillus* great production hosts are briefly introduced. The goal was to keep few short ideas per image.

More of the insights generated by this project could be included in the future. For example, some insights from the comparative genomic analysis. Prophage content could be presented as internal time bombs with potential implications in productivity, etc. These could be part of an infographic series to be distributed over social media for instance.

The last two graphics are conceptual cover proposals on how the series could be expanded to further *Bacillus* topics, in this case applications of bacterial pesticidal proteins for crop protection strategies and against nematocidal infections (See 6)



FROM ZERO TO HERO:

A BACTERIAL JOURNEY FROM SOIL TO INDUSTRY

Or the checkpoints a bacterium crosses to reach a bioreactor and become a microbial enzyme factory

WHY DO WE NEED ENZYMES?

Enzymes make our lives easier, they are proteins that speed up chemical reactions that otherwise would take too long to occur. Enzymes have many applications.

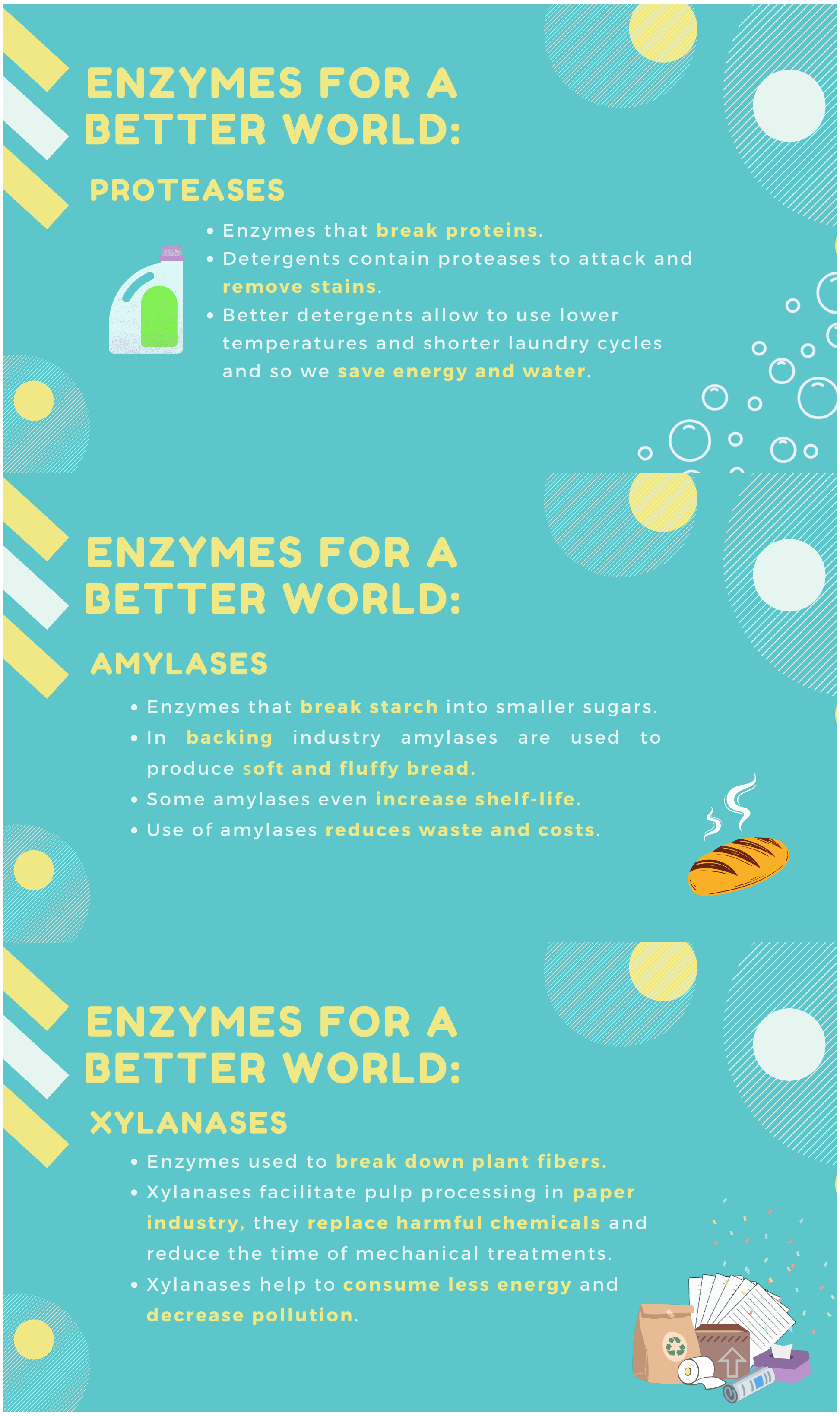
FOR EXAMPLE:

The enzyme β-galactosidase breaks down the lactose in milk so we can produce lactose-free milk.

ENZYMES FOR A BETTER WORLD

Enzymes are key for a more **sustainable economy**. We can replace toxic chemicals with biological processes. Enzymes are **biodegradable!**

Bioeconomy refers to the production and use of renewable sources in products, processes and services.



ENZYMES FOR A BETTER WORLD:

PROTEASES

- Enzymes that **break proteins**.
- Detergents contain proteases to attack and **remove stains**.
- Better detergents allow to use lower temperatures and shorter laundry cycles and so we **save energy and water**.

ENZYMES FOR A BETTER WORLD:

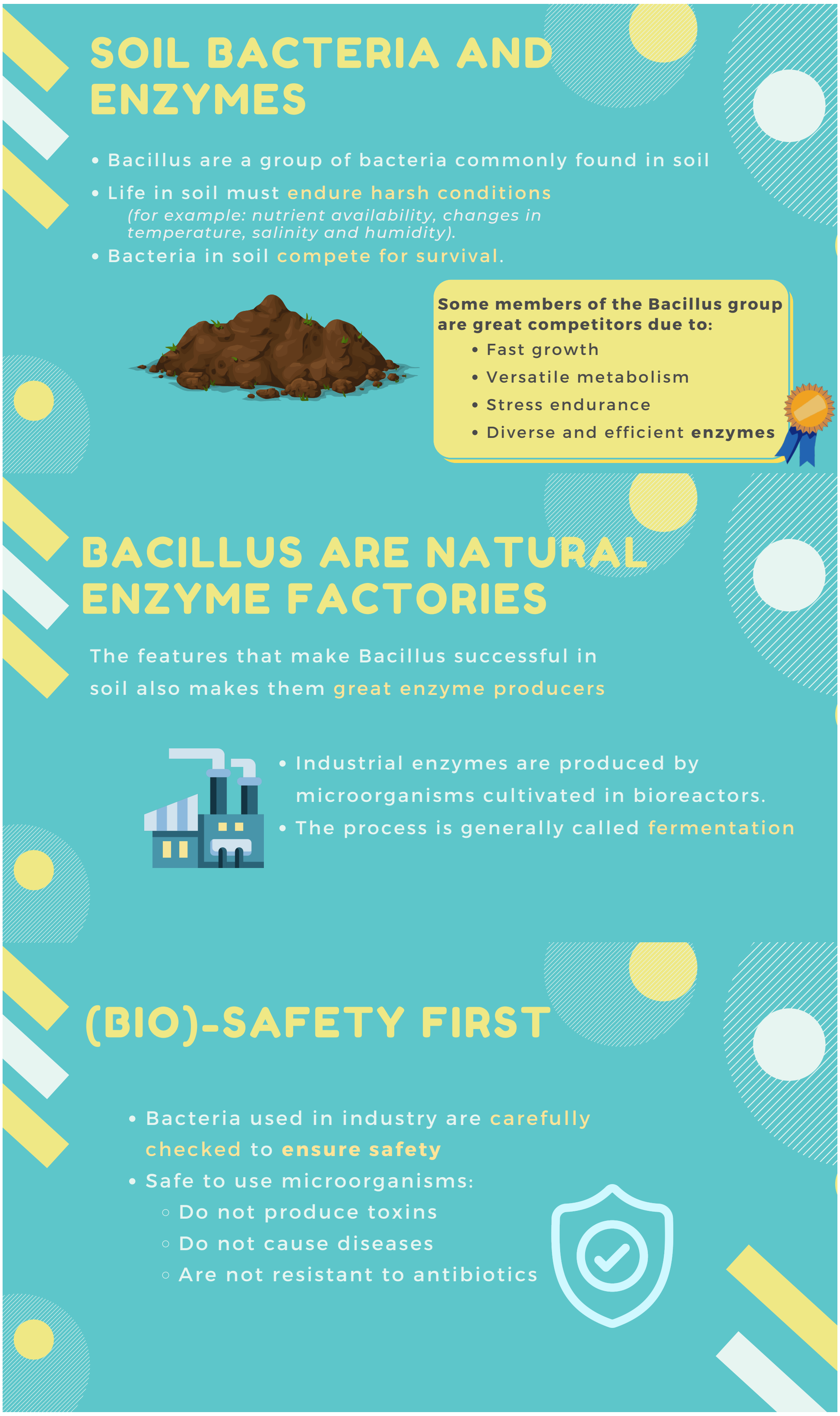
AMYLASES

- Enzymes that **break starch** into smaller sugars.
- In **backing** industry amylases are used to produce **soft and fluffy bread.**
- Some amylases even **increase shelf-life.**
- Use of amylases **reduces waste and costs**.

ENZYMES FOR A BETTER WORLD:

XYLANASES

- Enzymes used to **break down plant fibers.**
- Xylanases facilitate pulp processing in **paper industry**, they **replace harmful chemicals** and reduce the time of mechanical treatments.
- Xylanases help to **consume less energy** and **decrease pollution**.



SOIL BACTERIA AND ENZYMES

- Bacillus are a group of bacteria commonly found in soil
- Life in soil must endure harsh conditions
  (for example: nutrient availability, changes in temperature, salinity and humidity).
- Bacteria in soil compete for survival.

Some members of the Bacillus group are great competitors due to:

- Fast growth
- Versatile metabolism
- Stress endurance
- Diverse and efficient **enzymes**

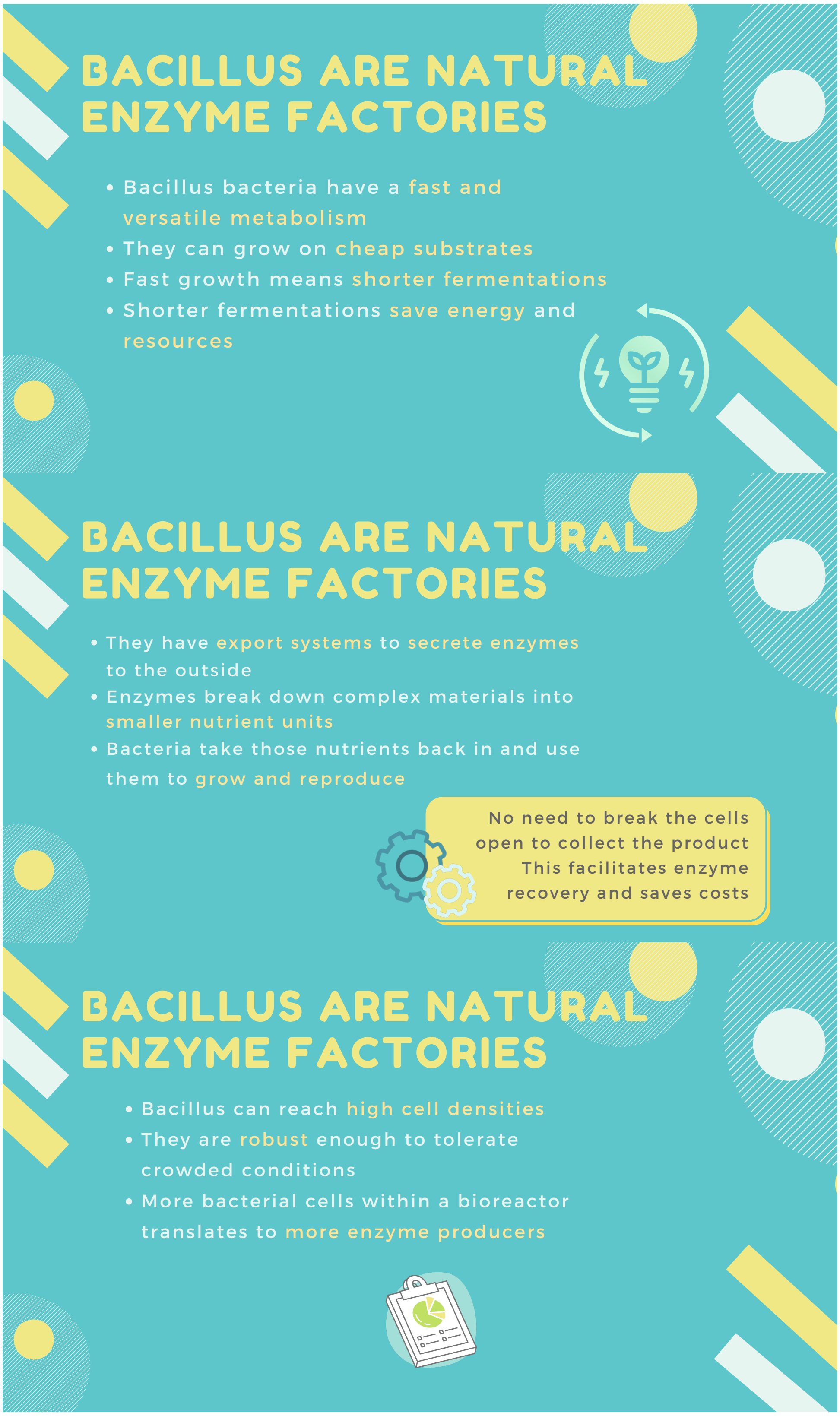
BACILLUS ARE NATURAL ENZYME FACTORIES

The features that make Bacillus successful in soil also makes them great enzyme producers

- Industrial enzymes are produced by microorganisms cultivated in bioreactors.
- The process is generally called fermentation

(BIO)-SAFETY FIRST

- Bacteria used in industry are carefully checked to **ensure safety**
- Safe to use microorganisms:
  - Do not produce toxins
  - Do not cause diseases
  - Are not resistant to antibiotics



BACILLUS ARE NATURAL ENZYME FACTORIES

- Bacillus bacteria have a fast and versatile metabolism
- They can grow on cheap substrates
- Fast growth means shorter fermentations
- Shorter fermentations save energy and resources

BACILLUS ARE NATURAL ENZYME FACTORIES

- They have export systems to secrete enzymes to the outside
- Enzymes break down complex materials into smaller nutrient units
- Bacteria take those nutrients back in and use them to grow and reproduce

No need to break the cells open to collect the product
This facilitates enzyme recovery and saves costs

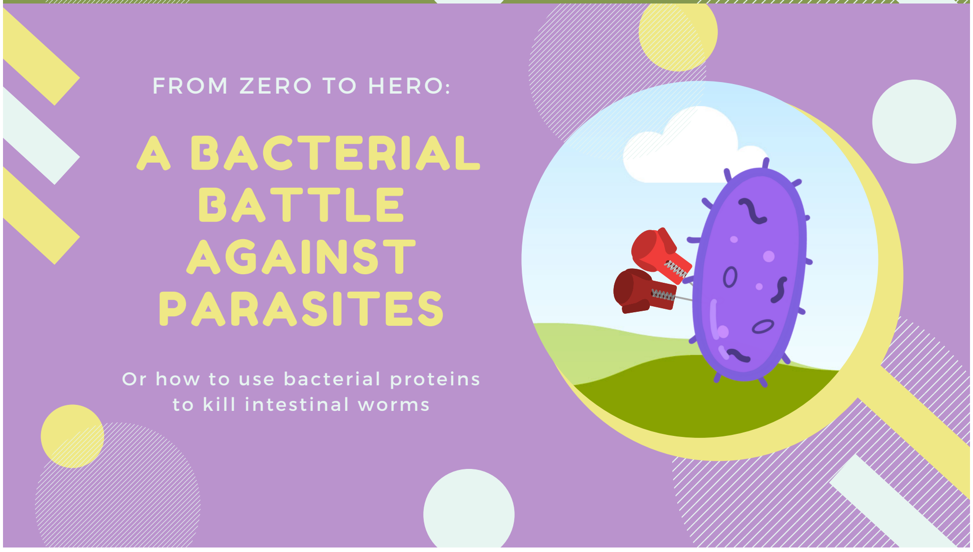
BACILLUS ARE NATURAL ENZYME FACTORIES

- Bacillus can reach high cell densities
- They are robust enough to tolerate crowded conditions
- More bacterial cells within a bioreactor translates to more enzyme producers

FROM ZERO TO HERO:

A BACTERIAL ALLIANCE TO PROTECT CROPS

Or how to find the right allies against agricultural pests



FROM ZERO TO HERO:

A BACTERIAL BATTLE AGAINST PARASITES

Or how to use bacterial proteins to kill intestinal worms

Chapter 6

Publications and manuscripts

This chapter includes the following publications and manuscripts:

- RNA of high yield, integrity and purity from industrial *Bacillus*, an improved method. *Status: under review*
- The complete genome of *Bacillus pumilus* MS32, insights on biotechnological production platforms. *Status: under review*
- IDOPS, a Profile HMM-Based Tool to Detect Pesticidal Sequences and Compare Their Genetic Context. *Status: published*
- Comparative Genomics of Chromosomally-Encoded Pesticidal Genes Reveals a Novel Prophage-Associated *cry* Cassette. *Status: in preparation*

Biology Methods & Protocols, 2023, pp. 1–9

doi: DOI HERE
Advance Access Publication Date: Day Month Year
Methods Article

METHODS ARTICLE

RNA of high yield, integrity and purity from industrial *Bacillus*, an improved method

Stefani Díaz Valerio ,[1] Mechthild Bömeke,[1] Anja Poehlein[1] and Heiko Liesegang[1]*

[1]Genomic and Applied Microbiology & Göttingen Genomics Laboratory, Georg-August University of Göttingen, Wilhelmsplatz 1, 37073, Göttingen, Germany

*Corresponding author. hlieseg@gwdg.de

FOR PUBLISHER ONLY Received on Date Month Year; revised on Date Month Year; accepted on Date Month Year

Abstract

Isolation of RNA of enough quality for downstream applications can turn burdensome, but remains critical for significant and reproducible results. Several commercial kits are available for reference organisms like *E. coli* K12 or *B. subtilis* 168 growing under optimal laboratory conditions. The situation for productive *Bacillus* strains growing at high cell densities in industrial fermentations is completely different. Our aim was to optimize an acid guanidinium thiocyanate–phenol–chloroform based protocol to return RNA of high yield, integrity and purity from *Bacillus* samples collected from small-scale fermentations resembling industrial conditions. The improved protocol includes modifications to overcome challenges such as: highly active RNases, incomplete cell lysis, organic extraction contaminants, high cell density media, poor precipitation efficiency and low RNA integrity. Phase Separating Gel facilitates this type of extractions, we describe how to prepare it using common laboratory supplies. By the implemented modifications, RNA samples from *B. pumilus* and *B. licheniformis* across all fermentation stages were suitable for generation of RNA-seq libraries. Library quality was further confirmed by determination of transcript integrity number (TIN). This protocol contributes to the study and characterization of biotechnologically relevant bacteria, which in turn facilitates the development and optimization of microbial cell factories.

Key words: RNA isolation, RNA-seq, RNA integrity number (RIN), RNA yield and purity, Transcript integrity number (TIN), Industrial *Bacillus*, fermentation

Introduction

Purified RNA of high quality is the starting point for several molecular biology methods, including reverse transcription quantitative real-time PCR (RT-qPCR), northern blot, microarray-analysis and next-generation RNA sequencing (RNA-seq) [56]. Advancements in these techniques enabled the study and growth of RNA biology, allowing gene expression analysis, transcriptomic profiling, discovery of novel non-coding RNAs, biomarker detection, etc. [59, 1, 21, 48]. These kind of investigations have not only highlighted crucial roles of RNA in cellular mechanisms, but also deepen our understanding on how an organism orchestrates genomic information into functional protein expression.

Selection of a RNA isolation approach, as the preceding step for many molecular techniques and assays, is critical. Considerations regarding sample collection, isolation conditions, RNA storage and compatibility with downstream applications are necessary to generate biologically significant results. Additionally, intrinsic features of the organism(s) of interest and the sample source are also variables to contemplate, sometimes in a case-dependent fashion. [33, 20, 37, 56, 15, 12, 18]. Generally, RNA isolation efforts are directed to retrieve RNA of the highest possible quality in terms of purity, yield and integrity. Organic extraction methods, column-based kits, and more recently, magnetic particle technologies are in the repertoire of RNA isolation alternatives [9, 50, 56].

Described by Chomczynski and Sacchi in 1987 [5, 6], and regarded as a gold standard [56], the acid guanidinium thiocyanate–phenol–chloroform (AGPC) approach is the method of choice in many laboratories. Compared to silica column methods, AGPC extractions are more versatile and robust, moreover they tend to retrieve higher yields of RNA [9, 61, 50, 62]. This is in part because commercial RNA isolation kits are developed and tested primarily with model organisms and in standard laboratory conditions, but fall short with more challenging strains and complex sample sources, such as the high cell density broth within a bioreactor.

Unlike domesticated laboratory strains, bacteria selected for their biotechnological performance, like many *Bacillus* strains, tend to be on the ”challenging to handle” side of the spectrum [18]. For example, some highly productive industrial strains of the species *B. pumilus* encode very active RNA degradation systems. In consequence, this leads to RNA of insufficient quality for RNA-seq when using standard kits and protocols.

© The Author 2023. Published by Oxford University Press. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com

The genus *Bacillus* is composed by Gram-positive, endospore-forming bacteria with a wide distribution in nature [35]. Members of the *B. subtilis* clade, such as *B. subtilis*, *B. licheniformis* and *B. pumilus* are highly used in biotechnology [49, 17]. These organisms excel as industrial workhorses due to its rapid growth-rates, robustness, and capacity to produce and secrete high amounts of extracellular enzymes, such as proteases, amylases and xylanases [8, 17, 19]. Additionally, they are considered as safe, there is a sound knowledge base and wide set of genetic manipulation techniques developed around them [49, 30].

However, as Gram-positive bacteria, *Bacillus* also posses a cell envelope composed of 50% - 80% of peptidoglycan, with around 10% of it being in association with teichoic acid, making these cells harder to lyse [51]. This constitutes a barrier to efficient high yield RNA isolation [18]. The difficulty to remove or inactivate ribonucleases (RNases) is another common obstacle to preserve RNA integrity and yield [37, 12, 15].

Despite the challenges, transcriptomic studies of *Bacillus* from conditions close to application scenarios like high cell density fermentations are crucial to understand the regulatory events that lead from substrate to product [17].

Current society goals regarding environment protection, energy supply and climate change, call for a transition into a circular bioeconomy framework [7]. Industrial biotechnology, which includes the development and optimization of microbial cell factories, is regarded as a key step in this transition [25, 43, 60].

By understanding the endogenous characteristics of a productive *Bacillus* strain, its performance and product yield can be optimized. This is achieved by integrating several ”omics” approaches, like genomics, proteomics, metabolomics and transcriptomics [17]. The last one of course, implies isolation of high quality RNA. RNA yield, purity and integrity are critical for the significance and reproducibility of such studies [50]. Therefore, we engaged in optimizing a RNA isolation protocol, with the focus on *Bacillus* of industrial interest and downstream RNA-seq applications.

We choose 3 species of *Bacillus* to run fermentations supporting high cell densities and resembling industrial conditions. The small-scale bioreactors were sampled at different time points of the whole process, and we used a modified phenol-based RNA isolation protocol highly implemented in literature as starting point [44]. However, samples from the bioreactors yielded RNA of poor quality, inadequate for RNA-seq experiments. Then we asked: what changes could result into increased RNA quality in terms of yield, integrity and purity? The improved protocol returned RNA suitable for RNA-seq applications. Moreover, it is consistent, reproducible, and adaptable, as we describe why and how the isolation steps could be modified to maximize RNA quality. This is relevant beyond our selected *Bacillus*, as the modifications here described can be applied in research of other biotechnologically relevant bacteria.

Materials and Methods

Preparations

Water for buffers and stock solutions was treated overnight with 0,1 % DEPC (diethyl-pyrocarbonate, Carl Roth, Karlsruhe, Germany) and autoclaved (20 minutes, 121 °C) before utilization. ”DNA-/DNase-/RNase-/PCR inhibitor free” low binding micro tubes (Sarstedt, Nümbrecht, Germany) were used during RNA extraction steps and for RNA storage. Sample processing, RNA handling, and library construction steps were done inside a RNA-dedicated laminar flow cabinet (AURA-PCR model from BIOAIR, Pero (MI), Italy). Before and after each work session the UV lamp of the cabinet was turned on for 20 minutes. Pipettes and work surfaces were further cleaned with RNase-ExitusPlus (AppliChem, Darmstadt, Germany). Analytical grade reagents and pipette tips with filter were used.

Buffers and solutions

The Killing-Buffer is composed of 20 mM Tris-HCl pH 7.5 (Carl Roth, Karlsruhe, Germany), 5 mM MgCl2 (Merk, Darmstadt, Germany) with 20 mM NaN3 (Merk, Darmstadt, Germany) and stored at 4 °C. The Lysis Buffer consists of 4 M Guanidine Thiocyanate (Carl Roth, Karlsruhe, Germany), 25 mM sodium acetate pH 5.2 (Carl Roth, Karlsruhe, Germany), and 0,5% N-lauroylsarcosinate (Sigma-Aldrich, Taufkirchen, Germany), this buffer can be stored at 4 °C for up to three months. Before RNA extraction the Lysis Buffer was placed in a water bath at 55 °C. A solution of sodium-acetate (Carl Roth, Karlsruhe, Germany) 3M, pH 5.2 was also prepared. Buffers and solutions were sterilized with 0.22 µm filters into new RNase-free tubes.

Phase Separating Gel

Phase Separating Gel (PSG) was prepared by mixing MOLYKOTE high vacuum grease (QUAX, Otzberg, Germany) with 12 % SiO_2 (particle size 0,5-10 µm, Sigma-Aldrich, Taufkirchen, Germany) in a 50 mL Falcon Tube with a metal spatula (it takes some time to homogenize due to its viscosity). Around 0,3-0,5g of the mixture were transferred to 2 ml low binding tubes, briefly centrifuged, autoclaved twice and UV treated (20 min) before usage. The use of common laboratory supplies such as high vacuum grease and SiO_2, to prepare customized PSG is based on previous work [40] and also described in biology blogs [26, 27].

Lysis Tubes

Glass beads of 0.1 mm and 0.2-0.5 mm (Carl Roth, Karlsruhe, Germany) were mixed in a 9:1 ratio and approx. 350 mg were transferred to 2 mL tubes with screw cap. The lysis tubes were autoclaved twice and stored. Shortly before the RNA extraction 75 µL of Buffer RLT (QIAGEN, Hilden, Germany) plus 1% BME (beta-mercaptoethanol, Carl Roth, Karlsruhe, Germany) were added and the lysis tubes were placed in a pre-cooled metal stand.

RNA isolation

Figure 1 depicts a schematic representation of the steps performed to isolate RNA from industrial *Bacillus* samples. To start, cell pellets were taken out from the -80 °C freezer, placed in a cold metal stand, resuspended in 150 µL of Buffer RLT (Qiagen, Hilden, Germany) plus 1% BME (beta-mercaptoethanol, Carl Roth, Karlsruhe, Germany) by pipetting up and down and transferred to the pre-cooled lysis tubes. Cell disruption was carried on a FastPrep-24 high-speed tissue homogenizer (MP Biomedicals, Eschwege, Germany) in three rounds of 40s with 1 minute ice incubation in between rounds, the equipment was set at 6.5 m/s. After cell disruption, 600 µL of pre-warmed Lysis Buffer (55 °C) were added and the sample was distributed in two 2 mL tubes containing the Phase Separating Gel (transfer of some glass beads does not affect the following steps).

One volume of PCI (Phenol-Chloroform-Isoamyl alcohol at 25:24:1, pH 4.5-5, Carl Roth, Karlsruhe, Germany) was added (e.g. 700 µL) to the samples, then the tubes were inverted 5-10 times and mixed at 700 rpm, for 3 minutes in a Thermomixer comfort device (Eppendorf, Hamburg, Germany) with occasional inverting (1-2 times). Then, the samples were centrifuged at 13000 rpm for 6 minutes at room temperature (MiniSpin Plus centrifuge, Eppendorf, Hamburg Germany). After centrifugation the Phase Separating Gel locates between the organic phase at the bottom of the tube and the upper aqueous phase. This facilitates the complete retrieval of the RNA-containing layer without carry-over or contamination from the interphase. The upper phase was transferred to a new 2 mL tube where one volume of PCI was added and the sample was inverted 5-10 times, mixed at 700 rpm for 3 minutes and centrifuged at 13000 rpm for 6 minutes. The upper aqueous phase was collected and transferred to a new 2 mL tube. Then, one volume of CI (Chloroform-Isoamyl alcohol at 24:1, Carl Roth, Karlsruhe, Germany) was added. Samples were inverted 5-10 times before mixing at 700 rpm for 3 minutes, followed by centrifugation at 13000 rpm for 6 minutes. Afterwards, the upper phase was carefully collected and transferred to a new 2 mL tube, to which 1/10 volume of sodium acetate (3M, pH 5.2, Carl Roth, Karlsruhe, Germany) was added followed by 1 µL of glycogen (Peqlab, Erlangen, Germany). The sample was gently vortexed for 3-5 seconds before addition of 2 volumes of ice cold absolute ethanol (Sigma-Aldrich, Taufkirchen, Germany), then it was mixed by vortexing and placed on the Thermomixer for 10 minutes, 700 rpm at 22 °C (Some precipitate might be already visible at this point). Finally, the samples were placed at the -20 °C freezer overnight.

RNA precipitation

After overnight incubation, the RNA precipitated from the solution and was collected by centrifugation before further ethanol washes to remove contaminants and ensure sample purity. A refrigerating centrifuge (Model 5417 R, Eppendorf, Hamburg, Germany) was pre-cooled to 4 °C and the samples were centrifuged at 13000 rpm, 20 minutes at 4 °C. The supernatant was carefully decanted to avoid dislodging the pellet and 1 mL of cold 70% ethanol was added. The tube was placed in the centrifuge so the pellet migrates to the opposite wall during centrifugation at 13000 rpm at 4 °C, 20 minutes. The supernatant was again decanted without disturbing the pellet and 1 mL of cold ethanol 70 % was added before repeating the centrifugation step, this time placing the tube so the pellet moves to the original position. The supernatant was carefully decanted and the remaining ethanol was collected by pipetting after a quick spin. To dry the pellet, the tubes were opened for 6 minutes, room temperature, inside the RNA-dedicated bench (over-drying of the pellet must be avoided). Nuclease-Free water (VWR International GmbH, Darmstadt, Germany) was pre-warmed at 52 °C and used to resuspend the RNA pellet. Finally the sample was incubated on ice for 3 hours before storage or further processing. Samples were initially resuspended in 50 µL of Nuclease-Free water and in case Qubit measurements indicated a concentration higher than 500 ng/µL, additional 50-150 µL of water were added accordingly.

Quality Assessment

The extracted RNA was evaluated for purity, yield and integrity using the following devices: a Nanodrop Spectrophotometer (ND-100, Thermo Fisher Scientific, USA), QuBit (catalog #Q32857, with the RNA Broad Range assay kit, Invitrogen) and a BioAnalyzer (Agilent 2100, Prokaryote Total RNA Nano assay, Agilent Technologies, Inc). For our application case RNA-seq libraries where generated, allowing for an additional quality control: determination of the transcript integrity number (TIN). TIN number was obtained with the tin.py function of the RSEQC package [58] (version 4.0.0) using as input alignment files and a genome reference in .bed format.

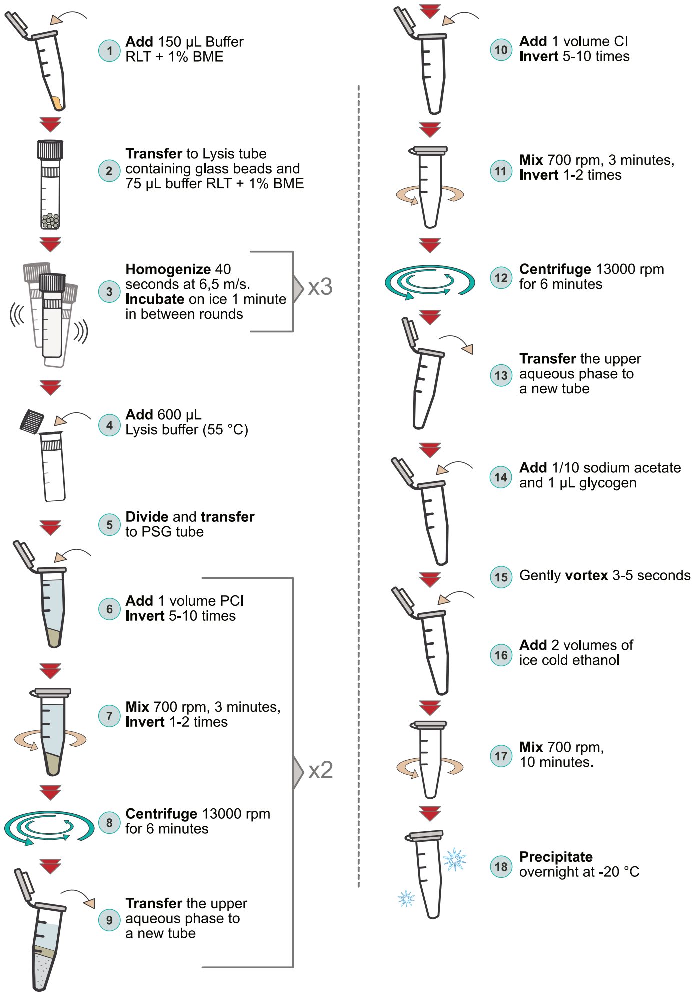

1 Add 150 µL Buffer RLT + 1% BME
2 Transfer to Lysis tube containing glass beads and 75 µL buffer RLT + 1% BME
3 Homogenize 40 seconds at 6,5 m/s. Incubate on ice 1 minute in between rounds
x3
4 Add 600 µL Lysis buffer (55 °C)
5 Divide and transfer to PSG tube
6 Add 1 volume PCI Invert 5-10 times
7 Mix 700 rpm, 3 minutes, Invert 1-2 times
x2
8 Centrifuge 13000 rpm for 6 minutes
9 Transfer the upper aqueous phase to a new tube
10 Add 1 volume CI Invert 5-10 times
11 Mix 700 rpm, 3 minutes, Invert 1-2 times
12 Centrifuge 13000 rpm for 6 minutes
13 Transfer the upper aqueous phase to a new tube
14 Add 1/10 sodium acetate and 1 µL glycogen
15 Gently vortex 3-5 seconds
16 Add 2 volumes of ice cold ethanol
17 Mix 700 rpm, 10 minutes.
18 Precipitate overnight at -20 °C

Fig. 1. Workflow for the RNA isolation from industrial *Bacillus* samples. Schematic representation of the optimized protocol here described. PCI: Phenol-chlorofom-isoamyl alcohol, CI: Chloroform-Isoamyl alcohol, BME: beta-mercaptoethanol

Application case

The improved RNA isolation protocol was applied to samples from three *Bacillus* species of industrial interest. First *B. subtilis* 168 Δ *yqfD* as model organism and well characterized cell factory, *B. licheniformis* MW3 Δ *yqfD* representing another established production platform and *B. pumilus* MS32 as emerging expression host. The selected strains are germination negative mutants, as it is commonly required to prevent contamination and additional sterilization costs within industrial facilities.

Fermentation conditions and sampling

B. pumilus MS32, *B. licheniformis* MW3 Δ *yqfD* [47] and *B. subtilis* 168 Δ *yqfD* were re-activated from lyophilized cultures and precultured twice during 16 and 6 hours, respectively, before inoculation of 0.5 L bioreactors. The pH within the fermenter started at 6,9 and afterwards controlled to 7.10 +/- 0.2 with 12.5% NH_3 or 12.5% H_2SO_4. Aeration was set to 0.5 vvm, the stirred speed was 1200 rpm, and the temperature 37 °C. Super rich fermentation media consisted of: 2% Yeast Extract, 2.5% Tryptone, 1% NaH_2PO_4 x2 H_2O, 1% Na_2HPO_4 x2 H_2O, 1% Saccharose and 0.5% Potato Extract Glucose Broth.

The fermentations were done at the Research and Development facilities of AB Enzymes GmbH, Darmstadt, Germany. Triplicate fermentation runs were done for each organism. Bioreactors were sampled at 4 time-points: 2.5, 4, 7 and 19 hours after inoculation. Fermentation broth was mixed with half volume of partially frozen Killing-Buffer (also referred as killing slurry) and 150 µL aliquots were made (250µL in the case of 2.5 hour samples). The microtubes were centrifuged 12 minutes, 13000 rpm, at 4 °C, quickly froze in liquid nitrogen and stored at -80 °C freezer until shipment on dry ice and RNA isolation as described above.

Library Preparation

After RNA isolation 24 samples were prepared for sequencing (12 *B. licheniformis* and 12 *B. pumilus*). DNA was digested with the Ambion™ TURBO DNA-free™ Kit (catalog #AM1907, Thermo Fisher Scientific, Waltham, USA), in presence of RiboLock RNase Inhibitor (catalog #EO0381, Thermo Fisher Scientific, Waltham, USA). Following digestion, the samples were purified with RNeasy MinElute Cleanup kit (catalog #74204, QIAGEN, Hilden, Germany) and a control PCR was done to ensure effective and complete DNA digestion. Afterwards, rRNA was enzymatically degraded with the Illumina Ribo-Zero Plus rRNA Depletion Kit (catalog #20040526, Illumina, Inc., San Diego, USA) using species-specific probes ordered from IDT (Integrated DNA Technologies, Inc., Iowa, USA). Then the samples were purified using Agencourt RNAClean XP beads (catalog #A63987, Beckman Coulter Life Sciences, Krefeld, Germany). The NEBNext® Ultra™ II Directional RNA Library Prep Kit for Illumina (catalog #E7760, New England Biolabs, Ipswich, MA, USA) with Unique Dual Mutiplex Oligos (catalog #E6440) was used to prepare cDNA libraries. Sequencing was done on ... (NovaSeq specs?).

Bioinformatic processing

FastP [4] (version 0.20.1) was used to asses the quality of the RNA-seq libraries, the detection of adapter sequences was enabled as well as the overrepresentated sequences analysis and the base-correction based on overlapping regions, the length filtering option was disabled. FastQ files with reads passing the FastP filter were passed to READemption ([13], version 1.0.10) for alignment and further analysis. Alignment files in BAM format were used as input to calculate the TIN number as described above.

Results and Discussion

Here we present an optimized RNA isolation protocol for bacterial samples of industrial interest. Figure 2 compares the improved protocol with a method traditionally used for *Bacillus*. The conventional method described by Petersohn et. al [44] was later on applied and adapted to multiple other studies [42, 31, 53, 24, 2, 63, 45]. The figure also summarizes the main advantages the optimized protocol, which will be discussed below.

Despite protocols and commercial kits for RNA isolation being available, no method can be universally applied [33]. Particular source materials and organisms still turn to be challenging when the goal is to produce RNA of high yield, quality and integrity. We faced such challenges when isolating RNA from non-standard *Bacillus* species sampled from bioreactors resembling industrial fermentation conditions. Figure 3 shows the yield and integrity obtained in early isolation rounds. An early protocol variation (disruption in Killing Buffer) successfully generated RNA of good quality for *B. subtilis* (RIN=9.7, 872 ng/µL), however, for the closely related *B. licheniformis* (RIN=7.1, 1063 ng/µL) and *B. pumilus*(RIN=5.4, 506 ng/µL) the RNA integrity was unsuitable for the downstream RNA-seq applications. As the samples were collected and processed in parallel, this pointed species-related differences behind the poor RNA integrity. Moreover it highlighted the need to further adapt and optimize the RNA isolation steps.

A possible explanation for the decreasing RNA integrity scores is endogenous ribonucleases being more abundant or active in *B. licheniformis*, and particularly in *B. pumilus*. In fact, some *B. pumilus* strains are investigated for its capacity to produce and secrete RNases with promising antiviral [10, 46] and anticancer [22] applications. A similar ribonuclease has also been described in *B. licheniformis* [41]. These soil bacteria might have evolved such RNases as an adaptation to use alternative phosphate sources under nutrient-limiting growth conditions, or even to use as bacteriocins against other bacterial populations [54, 41].

Cell disruption is a critical step in which RNA is highly vulnerable to the degrading activity of RNases. [36]. Lysis buffers often contain chaotrophic salts, like Guanidinium thiocyanate, to denature and prevent the nuclease activity. However, incomplete cell disruption can leave RNases unexposed to inactivating agents. The cell wall of Gram positive organisms imposes an efficient barrier to commonly used lysis buffers [55, 51], therefore, protocols often include additional disruption methods [18].

We selected bead beating in a high speed tissue homogenizer to ensure fast and complete cell disruption. The ice incubation steps are necessary to prevent sample overheating by mechanical forces, which has been reported as a potential drawback of long beating periods, leading to RNA degradation [32]. This lysis method has proven successful for difficult-to-crack organisms besides Gram-positive cells, such as filamentous fungi [32] and some environmental samples from complex matrices [33].

Other protocols recommend incubation with lyzozyme to digest the Gram-positive cell wall [55, 23], nevertheless this can result detrimental to the RNA integrity as endogenous RNases might become active during the incubation period [14]. As alternative, a micro-dismembrator device is conventionally used to disrupt Gram-positive cells, briefly, the sample is placed in a Teflon vessel, liquid nitrogen is added, and a tungsten carbide bead breaks the cells in cycles of 2 minutes [44]. Despite its effectiveness, the micro-dismembrator can process only one sample at a time and requires careful cleaning and assembly of the Teflon recipient in between samples. On the

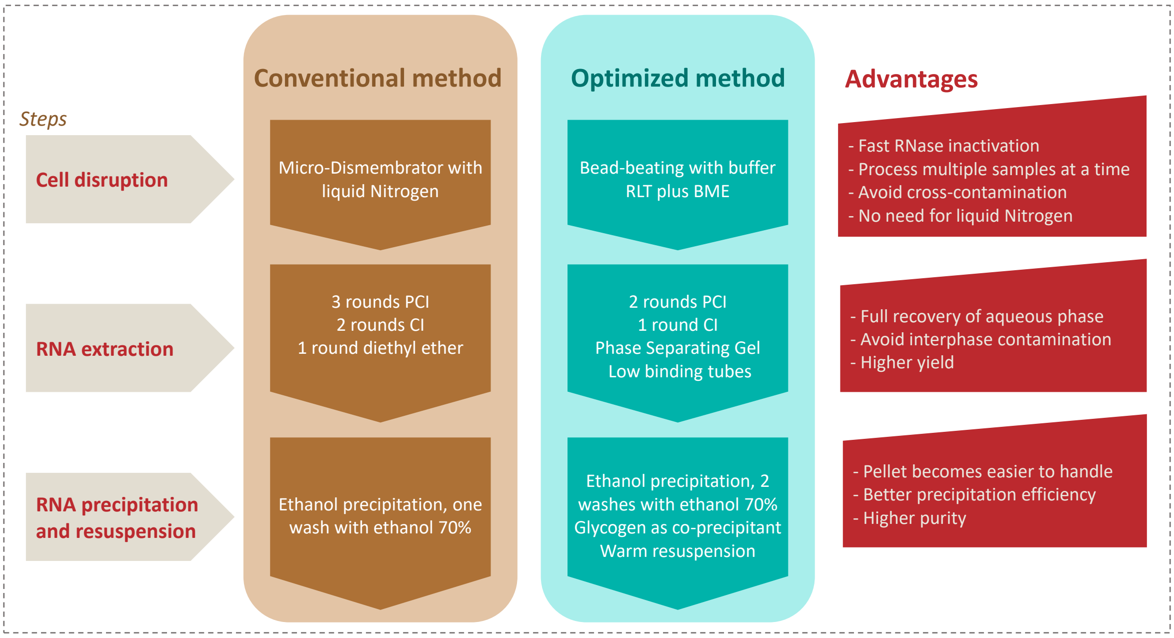

Conventional method
Optimized method
Advantages
Steps
Cell disruption
Micro-Dismembrator with liquid Nitrogen
Bead-beating with buffer RLT plus BME
- Fast RNase inactivation
- Process multiple samples at a time
- Avoid cross-contamination
- No need for liquid Nitrogen
RNA extraction
3 rounds PCI
2 rounds CI
1 round diethyl ether
2 rounds PCI
1 round CI
Phase Separating Gel
Low binding tubes
- Full recovery of aqueous phase
- Avoid interphase contamination
- Higher yield
RNA precipitation and resuspension
Ethanol precipitation, one wash with ethanol 70%
Ethanol precipitation, 2 washes with ethanol 70%
Glycogen as co-precipitant
Warm resuspension
- Pellet becomes easier to handle
- Better precipitation efficiency
- Higher purity

Fig. 2. Comparison between conventional and optimized methods for RNA isolation from *Bacillus* samples. This overview highlights the main differences and advantages of the method here presented. PCI: Phenol-chlorofom-isoamyl alcohol, CI: Chloroform-Isoamyl alcohol, BME: beta-mercaptoethanol.

contrary, by bead beating we can disrupt multiple samples simultaneously, avoid the need of liquid nitrogen, and the risk of cross-contamination between samples by sharing the recipient, making this alternative more convenient (Figure 2).

Pilot extractions, such as those depicted in Figure 3, were done by bead beating the cells in presence of killing buffer. Other tests included: disrupting in killing buffer supplemented with 1% beta-mercaptoethanol, and in lysis buffer (data not shown). Despite the last one having Guanidinium thiocyanate as inactivating agent, the N-lauroylsarcosinate produced foaming which prevented complete cell lysis. The best outcome was achieved by changing to buffer RLT supplemented by 1% beta-mercaptoethanol. Buffer RLT also contains guanidinium salts and addition of beta-mercaptoethanol further contributes to RNase inactivation by reducing the disulphide bonds in the tertiary structure of RNases [28, 38]. One disadvantage of beta-mercaptoethanol is its high toxicity and volatility, Dithiothreitol (DTT) has been recommended as a less toxic alternative [38].

Despite its advantages and popularity, phenol-chloroform based methods still require careful pipetting of the upper aqueous phase, which contains the RNA. The interphase is difficult to see and sensitive to agitation [33], imposing the hurdle to obtain as much of the upper phase without disturbing the interphase, while avoiding to take-in from the lower phase. If this step is not done properly, phenol contamination can affect subsequent enzymatic steps, and DNA leftovers will cause biases in quantification and downstream RNA-seq analysis due to the generation of reads derived from DNA instead of RNA.

To facilitate phase discrimination one can add 8-hydroxyquinoline to the phenol, which prevents its oxidation and turns it yellow, moreover it is a partial RNase inhibitor [12]. Commercial extraction solutions, such as TRIzol, are also colored in order to ease phase handling. Some protocols recommend to leave a small volume of the upper phase behind to avoid the risk of contamination, nevertheless this translates to reduced yield. We found the use of Phase Separating Gel as the best option to recover the RNA containing layer. It creates a physical barrier that blocks the organic phase and cell debris below, which not only avoids contamination but also allows for full recovery of the aqueous phase.

There are commercial options to the Phase Separating Gel, such as 5PRIME Phase Lock Gel (QuantaBio), PhaseMaker (Invitrogen) or Phase Divider Gel (Sigma-Aldrich), however they represent a higher cost and might not be suitable for every buffer system. For example, manufacturer indications recommend to use PhaseMaker tubes only in combination with TRIzol-based extractions. Here we describe a lesser known and cost-efficient alternative which can be easily customized to work with buffers and solutions of diverse densities [26, 27, 40]. For example, the density of the high vacuum grease was increased by addition of SiO_2, so it could properly migrate and locate at the interphase, this was due to the high salt content of the lysis buffer. Figure 4 shows examples of PSG with different SiO_2 concentrations, and how they migrated during testing rounds.

The chloroform-isoamyl alcohol (CI) step is recommended to remove traces of residual phenol, proteins, lipids and detergents that could remain associated with the RNA containing phase [33]. The amount of PCI and CI steps vary in literature, some protocols include an additional extraction with diethyl ether as well [39, 29]. Given the effectiveness of the PSG to separate the upper aqueous phase, the amount to PCI/CI rounds could be reduced, making this protocol more time-efficient. To avoid

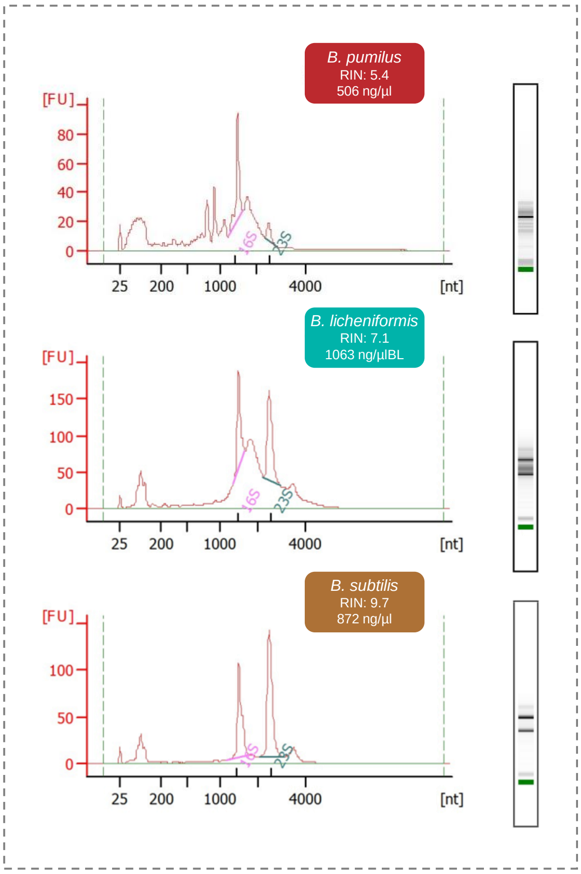


Fig. 3. RNA of high integrity is more difficult to obtain in non-model organisms. Bioanalyzer electropherograms show how an earlier version of the RNA isolation protocol successfully generated RNA of high integrity for *B. subtilis* but required optimization in the case of *B. licheniformis* and *B. pumilus*. The samples were processed in parallel with a protocol that used disruption in presence of killing buffer instead of buffer RLT supplemented with beta-mercaptoethanol.

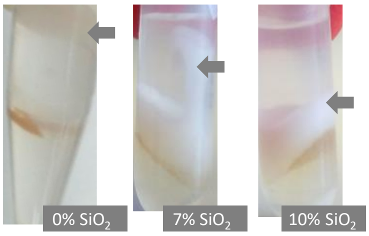


Fig. 4. Addition of SiO_2 allows the Phase Separating Gel (PSG) to stay at the interphase. Due to the high salt concentration of the lysis buffer, it was necessary to increase the density of the PSG so it could properly migrate and stay between the organic and the aqueous phases of the phenol-chloroform RNA extraction. Notice the cell debris being effectively separated from the RNA-containing layer. Arrows point to the PSG with different SiO_2 concentrations.

RNA left behind every time the sample is transferred to a new tube, low-binding tubes are preferred [3].

Regarding RNA precipitation, glycogen is added as a co-precipitant [34], this has two advantages. First, glycogen turns the pellet visible which facilitates sample handling and prevents accidental loss when decanting. Second, as an inert carrier, it allows to retrieve higher yields of RNA without interfering with downstream enzymatic applications, such as cDNA synthesis. To ensure sample purity, we preferred precipitation on ethanol. Salts are less soluble in isopropanol [16] and tend to precipitate together with the RNA, specially if the incubation is not done at room temperature. Another advantage of using ethanol is its volatility, making the drying of the pellet easier [15]. A systematic investigation on factors influencing nucleid acid precipitation found that overnight precipitation at -20 °C has better recovery when compared to 2 hour incubations, in the same investigation the authors reported higher yield for miRNA precipitated in ethanol in comparison to isopropanol [34].

The improved protocol was applied to the *B. pumilus* and *B. licheniformis* samples collected at different time points of small-scale fermentations. Table 1 depicts a characterization of the RNA obtained by the optimized isolation protocol. By applying the changes described above, it was possible to obtain RNA of high integrity, purity and yield from previously unsuccessfully isolated samples (Figure 3). The average yield for *B. pumilus* samples was 392.8 ng/µL (min = 240 ng/µL, max = 676 ng/µL), while for *B. licheniformis* it was 590.4 ng/µL (min = 266 ng/µL, max = 1521 ng/µL), which provided enough material for cDNA library construction. It is of notice that all samples had RNA integrity numbers (RIN) above 8, which is commonly accepted as high quality RNA suitable for RNA-seq experiments. For both organisms, RIN values were generally higher at earlier sampling points, nevertheless even from 19 hours samples, RIN was still higher than 8. The improved protocol successfully retrieves optimal RNA despite the accumulation of dead cells and by-products within the high cell density fermentation broth. These factors tend to complicate the RNA extraction procedure. Difficulties to obtain high quality RNA from *Bacillus* at late growth stages has been reported before [18]. Another common challenge is to obtain enough RNA from early time points, where there is a smaller ratio of bacterial cells in comparison to the highly concentrated complex media components. Here we prove that the measures adopted to maximize yield were efficient to isolate RNA even at 2.5 hours after inoculation of the bioreactor.

Beyond raw data quality assessment, there are other metrics that inform a researcher about the quality of an RNA-seq experiment before proceeding with downstream analysis. One of those metrics is the TIN value. First introduced in 2016, this metric stands for Transcript Integrity Number (TIN) and evaluates the mRNA integrity at sample and single transcript levels [57]. Calculation of TIN score is part of recommended guidelines for RNA-seq data assessment, which is of major relevance in order to avoid biases affecting downstream analysis such as identification of differentially expressed genes [52]. TIN score correlates to RIN values, and since it can be determined for each transcript, it is useful to correct biases such as those arising from differential RNA degradation [57]. RIN values have the disadvantage of being indirect estimations, relying in rRNA integrity and providing a number for the sample as a whole, which not necessarily reflects the state of all RNA species present in the sample. [11]. Therefore it is good practice to determine RIN values before committing to a RNA-seq experiment as a general indication for the sample, but ideally this should be complemented by calculation of TIN scores.

Table 1. Feature summary for the RNA isolated from *B. licheniformis* MW3 Δ *yqfD* and *B. pumilus* MS32. Time points represent sampling time after bioreactor inoculation, values correspond to the mean (n=3) ± standard deviation. Absorbance ratios were determined by Nanodrop spectophotometer, RIN values as reported by BioAnalyzer and TIN scores as calculated by RSEQC package. RIN:RNA Integrity Number, TIN: Transcript Integrity Number.

	B. pumilus				*B. licheniformis*			
	2.5h	4h	7 h	19h	2.5h	4h	7h	19h
A260/280	2.05 ± 0.04	2.05 ± 0.01	2.07 ± 0.01	2.00 ± 0.03	2.06 ± 0.02	2.06 ± 0.02	2.05 ± 0.03	2.07 ± 0.07
A260/230	2.11 ± 0.25	2.17 ± 0.26	2.09 ± 0.20	1.42 ± 0.22	1.64 ± 0.59	1.80 ± 0.60	2.09 ± 0.28	1.90 ± 0.35
RIN	9.63 ± 0.06	9.07 ± 0.21	8.83 ± 0.15	8.50 ± 0.44	9.73 ± 0.15	9.50 ± 0.26	8.73 ± 0.84	8.43 ± 0.15
TIN	87.80 ± 0.30	88.29 ± 0.16	86.63 ± 1.43	87.62 ± 0.59	87.27 ± 0.64	85.28 ± 0.30	86.28 ± 0.92	82.97 ± 5.11

Timing is the main limitation of this optimized protocol, particularly when compared to commercial kits and purification columns. Preparation of the Phase Separating Gel, overnight incubations and additional washing steps consume time. Use of hazardous substances such as phenol and beta-mercaptoethanol imply additional protective measures and adequate disposal management. Nevertheless, these are compensated by the results and the advantages of the changes described. These modifications are worth of consideration in situations where other methods failed to deliver optimal RNA quality and samples are precious. Like in our application case, were the sample amount per time point was limited and repetition of the whole fermentation experiments would have been too costly.

This optimized protocol overcomes the following challenges: 1-) Working with non-standard, difficult to lyse organisms with highly active RNases, specifically *Bacillus* of biotechnological relevance; 2-)Recovery of RNA of high purity out of complex media such as a fermentation broth; and 3-) Retrieving RNA of high integrity, and yield from early and late sampling points of a fermentation, each of which entail their own complications. Moreover, we described how to prepare and customize a Phase Separating Gel, which improves the isolation and is more cost-efficient than the commercial alternatives. We believe this protocol, or at least some of its modifications, could be applied in other similar experimental setups.

A great amount of time, dedication and resources were invested in the optimization of this protocol. Similar research focused on organisms of industrial interest might benefit from the modifications here described. By giving a more direct path to RNA of optimal quality, downstream analysis can take place earlier. Ultimately, this opens the door to a better understanding of the behavior of bacteria under industrial fermentation conditions, which in turn favors the so needed development, application and optimization of microbial cell factories.

Acknowledgements

We express our gratitude to Birgit Pfeiffer for valuable discussion during the early extraction experiments. To the Göttingen Genomics Laboratory (G2L), in particular to Melanie Heinemann and Sarah Schüßler for their technical assistance. We are also thankful to the AB Enzymes team, without their support, fermentation experiments and sampling would not have been possible. We acknowledge support by the Open Access Publication Funds of the University of Göttingen. This work was supported by the Federal Ministry of Education and Research (BMBF), Germany, grant ID: Fkz. 031B0573 and AB Enzymes GmbH. This work used the Scientific Compute Cluster at GWDG, the joint data center of Max Planck Society for the Advancement of Science (MPG) and University of Göttingen.

Competing interests

No competing interest is declared.

Author contributions statement

SDV performed experiments, analyzed data, developed new protocol, wrote manuscript, AP provided protocols, coordinated sequencing efforts, analyzed data, MB provided protocols, transferred RNA and cDNA library preparation techniques, performed experiments, HL designed study, analyzed data, wrote manuscript. All authors read, checked and agreed to the manuscript.

References

1. Lars Barquist and Jörg Vogel. Accelerating discovery and functional analysis of small rnas with new technologies. *Annual review of genetics*, 49:367–394, 2015.
2. Ina Budde, Leif Steil, Christian Scharf, Uwe Völker, and Erhard Bremer. Adaptation of bacillus subtilis to growth at low temperature: a combined transcriptomic and proteomic appraisal. *Microbiology*, 152(3):831–853, 2006.
3. Kasandra L Burgos and Kendall Van Keuren-Jensen. Rna isolation for small rna next-generation sequencing from acellular biofluids. In *RNA Mapping*, pages 83–92. Springer, 2014.
4. Shifu Chen, Yanqing Zhou, Yaru Chen, and Jia Gu. fastp: an ultra-fast all-in-one fastq preprocessor. *Bioinformatics*, 34(17):i884–i890, 2018.
5. Piotr Chomczynski and Nicoletta Sacchi. Single-step method of rna isolation by acid guanidinium thiocyanate-phenol-chloroform extraction. *Analytical biochemistry*, 162(1):156–159, 1987.
6. Piotr Chomczynski and Nicoletta Sacchi. The single-step method of rna isolation by acid guanidinium thiocyanate–phenol–chloroform extraction: twenty-something years on. *Nature protocols*, 1(2):581–585, 2006.
7. European Commission, Directorate-General for Research, and Innovation. *A sustainable bioeconomy for Europe : strengthening the connection between economy, society and the environment : updated bioeconomy strategy.* Publications Office, 2018.
8. Wenjing Cui, Laichuang Han, Feiya Suo, Zhongmei Liu, Li Zhou, and Zhemin Zhou. Exploitation of bacillus subtilis as a robust workhorse for production of heterologous

proteins and beyond. *World Journal of Microbiology and Biotechnology*, 34(10):1–19, 2018.

9. Yuri F Drygin, Konstantin O Butenko, and Tatiana V Gasanova. Environmentally friendly method of rna isolation. *Analytical Biochemistry*, 620:114113, 2021.
10. MA Efimova, R Shah Mahmud, PV Zelenikhin, MI Sabirova, AI Kolpakov, and ON Ilinskaya. Exogenous bacillus pumilus rnase (binase) suppresses the reproduction of reovirus serotype 1. *Molecular Biology*, 51(1):96–101, 2017.
11. Anna Esteve-Codina. Rna-seq data analysis, applications and challenges. *Comprehensive Analytical Chemistry*, 82:71–106, 2018.
12. RE Farrell Jr. Resilient ribonucleases. *RNA Methodologies: A Laboratory Guide for Isolation and Characterization, 4th ed.; Academic Press: Cambridge, MA, USA*, 2010.
13. Konrad U Förstner, Jörg Vogel, and Cynthia M Sharma. Reademption—a tool for the computational analysis of deep-sequencing–based transcriptome data. *Bioinformatics*, 30(23):3421–3423, 2014.
14. Patricia E Garrett, Feng Tao, Nathan Lawrence, Jay Ji, Richard T Schumacher, and Mark M Manak. Tired of the same old grind in the new genomics and proteomics era? *Targets*, 1(5):156–162, 2002.
15. Alan S Gerstein. *Molecular biology problem solver: a laboratory guide*. John Wiley & Sons, 2004.
16. Michael R Green and Joseph Sambrook. Precipitation of dna with isopropanol. *Cold Spring Harbor Protocols*, 2017(8):pdb–prot093385, 2017.
17. Yang Gu, Xianhao Xu, Yaokang Wu, Tengfei Niu, Yanfeng Liu, Jianghua Li, Guocheng Du, and Long Liu. Advances and prospects of bacillus subtilis cellular factories: from rational design to industrial applications. *Metabolic engineering*, 50:109–121, 2018.
18. Jean-Sebastien Guez, François Coutte, Anne-Sophie Drucbert, Nour-Eddine Chihib, Pierre-Marie Danzé, and Philippe Jacques. Resistance of the cell wall to degradation is a critical parameter for isolation of high quality rna from natural isolates of bacillus subtilis. *Archives of microbiology*, 191(8):669–673, 2009.
19. Colin R Harwood and Rocky Cranenburgh. Bacillus protein secretion: an unfolding story. *Trends in microbiology*, 16(2):73–79, 2008.
20. Rajandas Heera, Parimannan Sivachandran, Suresh V Chinni, Joanne Mason, Larry Croft, Manickam Ravichandran, and Lee Su Yin. Efficient extraction of small and large rnas in bacteria for excellent total rna sequencing and comprehensive transcriptome analysis. *BMC research notes*, 8(1):1–11, 2015.
21. Mingye Hong, Shuang Tao, Ling Zhang, Li-Ting Diao, Xuanmei Huang, Shaohui Huang, Shu-Juan Xie, Zhen-Dong Xiao, and Hua Zhang. Rna sequencing: new technologies and applications in cancer research. *Journal of hematology & oncology*, 13(1):1–16, 2020.
22. Olga N Ilinskaya, Indrabahadur Singh, Elena Dudkina, Vera Ulyanova, Airat Kayumov, and Guillermo Barreto. Direct inhibition of oncogenic kras by bacillus pumilus ribonuclease (binase). *Biochimica et Biophysica Acta (BBA)-Molecular Cell Research*, 1863(7):1559–1567, 2016.
23. Phetcharat Jaiaue, Piroonporn Srimongkol, Sitanan Thitiprasert, Somboon Tanasupawat, Benjamas Cheirsilp, Suttichai Assabumrungrat, and Nuttha Thongchul. A modified approach for high-quality rna extraction of spore-forming bacillus subtilis at varied physiological stages. *Molecular Biology Reports*, 48(10):6757–6768, 2021.
24. Tanja Kaan, Georg Homuth, Ulrike Mäder, Julia Bandow, and Thomas Schweder. Genome-wide transcriptional profiling of the bacillus subtilis cold-shock response. *Microbiology*, 148(11):3441–3455, 2002.
25. Manfred Kircher. Bioeconomy of microorganisms. In *The bioeconomy system*, pages 85–103. Springer, 2022.
26. Alex Klenov. Diy phase separating gel: Clean and cheap!, Dec 2013.
27. Alex Klenov. Dense phase separating gel / homemade trizol combo, Nov 2018.
28. Tony A Klink, Kenneth J Woycechowsky, Kimberly M Taylor, and Ronald T Raines. Contribution of disulfide bonds to the conformational stability and catalytic activity of ribonuclease a. *European Journal of Biochemistry*, 267(2):566–572, 2000.
29. Stefan Krebs, Marlis Fischaleck, and Helmut Blum. A simple and loss-free method to remove trizol contaminations from minute rna samples. *Analytical biochemistry*, 387(1):136–138, 2009.
30. Tobias Küppers, Victoria Steffen, Hendrik Hellmuth, Timothy O'Connell, Johannes Bongaerts, Karl-Heinz Maurer, and Wolfgang Wiechert. Developing a new production host from a blueprint: Bacillus pumilus as an industrial enzyme producer. *Microbial cell factories*, 13(1):1–11, 2014.
31. Lars Ingo Ole Leichert, Christian Scharf, and Michael Hecker. Global characterization of disulfide stress in bacillus subtilis. *Journal of Bacteriology*, 185(6):1967–1975, 2003.
32. Gonçalo M Leite, Naresh Magan, and Ángel Medina. Comparison of different bead-beating rna extraction strategies: an optimized method for filamentous fungi. *Journal of microbiological methods*, 88(3):413–418, 2012.
33. Mark A Lever, Andrea Torti, Philip Eickenbusch, Alexander B Michaud, Tina Šantl-Temkiv, and Bo Barker Jørgensen. A modular method for the extraction of dna and rna, and the separation of dna pools from diverse environmental sample types. *Frontiers in Microbiology*, 6:476, 2015.
34. Yalin Li, Suxiang Chen, Nan Liu, Lixia Ma, Tao Wang, Rakesh N Veedu, Tao Li, Fengqiu Zhang, Huiyue Zhou, Xiang Cheng, et al. A systematic investigation of key factors of nucleic acid precipitation toward optimized dna/rna isolation. *BioTechniques*, 68(4):191–199, 2020.
35. Niall A. Logan and Paul De Vos. *Bacillus*, pages 1–163. John Wiley & Sons, Ltd, 2015.
36. Lori A Martin, Tiffany J Smith, Dawn Obermoeller, Brian Bruner, Martin Kracklauer, and Subramanian Dhamaraj. Rna purification. *Molecular Biology Problem Solver: A Laboratory Guide*, pages 197–224, 2001.
37. Ralf Moeller, Gerda Horneck, Petra Rettberg, H-J Mollenkopf, E Stackebrandt, and WL Nicholson. A method for extracting rna from dormant and germinating bacillus subtilis strain 168 endospores. *Current microbiology*, 53(3):227–231, 2006.
38. Kathleen Mommaerts, Ignacio Sanchez, Fay Betsou, and William Mathieson. Replacing β-mercaptoethanol in rna extractions. *Analytical biochemistry*, 479:51–53, 2015.
39. David Moore and Dennis Dowhan. Purification and concentration of dna from aqueous solutions. *Current protocols in molecular biology*, 59(1):2–1, 2002.
40. Tapas Mukhopadhyay and Jack A Roth. Silicone lubricant enhances recovery of nucleic acids after phenol-chloroform

extraction. *Nucleic acids research*, 21(3):781, 1993.

41. Thanh Trung Nguyen, Minh Hung Nguyen, Huy Thuan Nguyen, Hoang Anh Nguyen, Thi Hoi Le, Thomas Schweder, and Britta Jurgen. A phosphate starvation-inducible ribonuclease of bacillus licheniformis. *Journal of Microbiology and Biotechnology*, 26(8):1464–1472, 2016.
42. Pierre Nicolas, Ulrike Mäder, Etienne Dervyn, Tatiana Rochat, Aurélie Leduc, Nathalie Pigeonneau, Elena Bidnenko, Elodie Marchadier, Mark Hoebeke, Stéphane Aymerich, et al. Condition-dependent transcriptome reveals high-level regulatory architecture in bacillus subtilis. *Science*, 335(6072):1103–1106, 2012.
43. Alessandro Pellis, Sara Cantone, Cynthia Ebert, and Lucia Gardossi. Evolving biocatalysis to meet bioeconomy challenges and opportunities. *New biotechnology*, 40:154–169, 2018.
44. Anja Petersohn, Matthias Brigulla, Stefan Haas, Jörg D Hoheisel, Uwe Völker, and Michael Hecker. Global analysis of the general stress response of bacillus subtilis. *Journal of bacteriology*, 183(19):5617–5631, 2001.
45. Philipp F Popp, Alhosna Benjdia, Henrik Strahl, Olivier Berteau, and Thorsten Mascher. The epipeptide yydf intrinsically triggers the cell envelope stress response of bacillus subtilis and causes severe membrane perturbations. *Frontiers in microbiology*, 11:151, 2020.
46. Daria S Pudova, Anna A Toymentseva, Natalia E Gogoleva, Elena I Shagimardanova, Ayslu M Mardanova, and Margarita R Sharipova. Comparative genome analysis of two bacillus pumilus strains producing high level of extracellular hydrolases. *Genes*, 13(3):409, 2022.
47. Michael Rachinger, Melanie Bauch, Axel Strittmatter, Johannes Bongaerts, Stefan Evers, Karl-Heinz Maurer, Rolf Daniel, Wolfgang Liebl, Heiko Liesegang, and Armin Ehrenreich. Size unlimited markerless deletions by a transconjugative plasmid-system in bacillus licheniformis. *Journal of biotechnology*, 167(4):365–369, 2013.
48. Antoine-Emmanuel Saliba, Sara C Santos, and Jörg Vogel. New rna-seq approaches for the study of bacterial pathogens. *Current opinion in microbiology*, 35:78–87, 2017.
49. Marcus Schallmey, Ajay Singh, and Owen P Ward. Developments in the use of bacillus species for industrial production. *Canadian journal of microbiology*, 50(1):1–17, 2004.
50. Amanda N Scholes and Jeffrey A Lewis. Comparison of rna isolation methods on rna-seq: implications for differential expression and meta-analyses. *BMC genomics*, 21(1):1–9, 2020.
51. Mohammed Shehadul Islam, Aditya Aryasomayajula, and Ponnambalam Ravi Selvaganapathy. A review on macroscale and microscale cell lysis methods. *Micromachines*, 8(3):83, 2017.
52. Keunhong Son, Sungryul Yu, Wonseok Shin, Kyudong Han, and Keunsoo Kang. A simple guideline to assess the characteristics of rna-seq data. *BioMed Research International*, 2018, 2018.
53. Leif Steil, Tamara Hoffmann, Ina Budde, Uwe Völker, and Erhard Bremer. Genome-wide transcriptional profiling analysis of adaptation of bacillus subtilis to high salinity. *Journal of bacteriology*, 185(21):6358–6370, 2003.
54. Vera Ulyanova, Valentina Vershinina, and Olga Ilinskaya. Barnase and binase: twins with distinct fates. *The FEBS Journal*, 278(19):3633–3643, 2011.
55. Eber Villa-Rodríguez, Cuauhtemoc Ibarra-Gámez, and Sergio de Los Santos-Villalobos. Extraction of high-quality rna from bacillus subtilis with a lysozyme pre-treatment followed by the trizol method. *Journal of microbiological methods*, 147:14–16, 2018.
56. I Vomelova, Z Vaníčková, and A Šedo. Technical note methods of rna purification. all ways (should) lead to rome. *Folia Biologica (Praha)*, 55:243–251, 2009.
57. Liguo Wang, Jinfu Nie, Hugues Sicotte, Ying Li, Jeanette E Eckel-Passow, Surendra Dasari, Peter T Vedell, Poulami Barman, Liewei Wang, Richard Weinshiboum, et al. Measure transcript integrity using rna-seq data. *BMC bioinformatics*, 17(1):1–16, 2016.
58. Liguo Wang, Shengqin Wang, and Wei Li. Rseqc: quality control of rna-seq experiments. *Bioinformatics*, 28(16):2184–2185, 2012.
59. Alexander J Westermann and Jörg Vogel. Cross-species rna-seq for deciphering host–microbe interactions. *Nature Reviews Genetics*, 22(6):361–378, 2021.
60. Sven Wydra. Value chains for industrial biotechnology in the bioeconomy-innovation system analysis. *Sustainability*, 11(8):2435, 2019.
61. Xueyu Xiang, Diwen Qiu, Richard D Hegele, and Wan C Tan. Comparison of different methods of total rna extraction for viral detection in sputum. *Journal of virological methods*, 94(1-2):129–135, 2001.
62. Baltasar Zepeda and Julian C Verdonk. Rna extraction from plant tissue with homemade acid guanidinium thiocyanate phenol chloroform (agpc). *Current Protocols*, 2(1):e351, 2022.
63. Jessica C Zweers, Pierre Nicolas, Thomas Wiegert, Jan Maarten van Dijl, and Emma L Denham. Definition of the σw regulon of bacillus subtilis in the absence of stress. *PloS one*, 7(11):e48471, 2012.

mSystems Submission Template mSystems Submission Template mSystems Submission Template mSystems Submission Template mSystems Submission Template mSystems Submission Template

The complete genome of *Bacillus pumilus* MS32, insights on biotechnological production platforms

Stefani Diaz Valerio,[a] Michael Seefried,[b] Ruth Schwerdtfeger,[b] Sonja Volland, [a] Anja Poehlein,[a] Rolf Daniel,[a] and Heiko Liesegang,[a,*]

Georg-August University of Göttingen, Institute of Microbiology and Genetics, Genomic and Applied Microbiology & Göttingen Genomics Laboratory, Göttingen, Germany[a]; AB Enzymes GmbH, Feldbergstrasse 78, D-64293 Darmstadt, Germany[b];

Compiled December 20, 2022
This is a draft manuscript, pre-submission

Address correspondence to Heiko Liesegang, hlieseg@gwdg.de.

* Institute for Microbiology and Genetics, Grisebachstr. 8, D-37077 Göttingen, Germany.

ABSTRACT Bacteria from the *Bacillus* genus are widely used in industrial protein production. *B. pumilus* is a new emerging expression platform for enzymes. The optimization of a production strain strongly benefits from genomic data, therefore we present the complete annotated genome of *Bacillus pumilus* MS32.

GENOME ANNOUNCEMENT

Bacillus pumilus is a ubiquitously distributed Gram-positive endospore-forming bacterium (1). The strain MS32 rapidly reaches high cell densities that produce and secrete large amounts of enzymes, a feature shared with other members of the *Bacillus* genus. *B. pumilus* strains have been identified as production hosts of valuable compounds like surfactins (2), lipases (3), laccases (4), chitinases (5), keratinases (6), collagenases (7) and other proteases (8, 9), attracting the interest of food, feed, detergent, leather and paper industries (10). *B. pumilus* is also investigated for agronomic (11, 12), probiotic (13, 14) and antitumor (15) applications.

B. pumilus MS32 was isolated from soil (North Rhine-Westphalia, Germany). After cultivation on LB media, total DNA was extracted with the MasterPure complete DNA and RNA purification kit following manufacturer recommendations (Epicentre, Madison, WI). Libraries were prepared from 1.5 μg of high-molecular-weight DNA using the Ligation Sequencing SQK-LSK109 and the Native Barcode Expansion (EXP-NBD114-Barcode5) kits. Sequencing was done with a MinION Mk1B device, SpotON flow cell R9.4.1, and MinKNOW 21.06.0 as recommended by the manufacturer (Oxford Nanopore Technologies). Guppy HAC 5.0.16 was used for demultiplexing and base calling. NanoFilt 2.8.0 (16) was used for trimming (13bp at each end) and filtering (average quality score > 12, length > 1000bp). Illumina short reads (provided by AB Enzymes) were processed with Fastp (17) by trimming (6bp at the front) and clipping (sliding window size=4, mean quality ≥ 20). A hybrid *de novo* assembly of processed long (262 001) and short (3 795 211) reads was done by Unicycler 0.4.9 (18) with conservative mode and default parameters. Annotation was done by PGAP 2021-07-01.build5508 (19). Average nucleotide identity (ANI) between the strain MS32 and complete *B. pumilus* genomes at NCBI was calculated with PyANI 0.2.11 (20). The genome was further characterized with antiSMASH 6.0.1 (21), PhiSpy 4.2.19 + VOG database (vog210, http://vogdb.org/) (22), ISEScan 1.7.2.3 (23) and CRISPRCasFinder 4.2.20 (24).

The genome of *B. pumilus* MS32 consists of a single 3,824,664bp chromosome, with a G+C content of 41.6% and encodes 3,712 proteins. ANI values with *B. pumilus* strains NCTC10337 (97.66%) and BIM B-171(97.50%) confirm MS32 as a member of the

ms Submission Template mSystems Submission Template mSystems Submission Template mSystems Submission Template mSystems Submission Template mSystems Submission Te

species. MS32 encodes all genes necessary for genetic competence. Genome analysis of MS32 revealed: 2 prophage regions, 37 IS elements, and one CRISPR array with no *cas* gene. antiSMASH indicates 13 putative secondary metabolite gene clusters, including those for lichenysin, plantazolicin and bacillibactin production. A gene cluster for a lanthipeptide of the less known class III seems unique to MS32. This characterization points to particular features and optimization targets to further develop *B. pumilus* MS32 as an industrial production platform.

Data Availability: The genome of *B. pumilus* MS32 is available at DDBJ/EA/GenBank (accession number CP092829). Raw reads are at the NCBI sequence read archive: SRR18190495 (Illumina) and SRR18190496 (Oxford Nanopore). *B. pumilus* MS32 is part of Westerdijk Institute's public collection, strain number CBS 140336.

ACKNOWLEDGMENTS

We acknowledge support by the Open Access Publication Funds of the University of Göttingen. This work was supported by the Federal Ministry of Education and Research (BMBF), Germany, grant ID: Fkz. 031B0573 and AB Enzymes GmbH. This work used the Scientific Compute Cluster at GWDG, the joint data center of Max Planck Society for the Advancement of Science (MPG) and University of Göttingen. We thank Melanie Heinemann and Mechthild Bömeke for technical support.

REFERENCES

1. **Liu Y, Lai Q, Dong C, Sun F, Wang L, Li G, Shao Z**. 2013. Phylogenetic diversity of the Bacillus pumilus group and the marine ecotype revealed by multilocus sequence analysis. PloS one 8 (11):e80097.
2. **Slivinski CT, Mallmann E, de Araújo JM, Mitchell DA, Krieger N**. 2012. Production of surfactin by Bacillus pumilus UFPEDA 448 in solid-state fermentation using a medium based on okara with sugarcane bagasse as a bulking agent. Process biochemistry 47 (12):1848–1855.
3. **Kim HK, Choi HJ, Kim MH, Sohn CB, Oh TK**. 2002. Expression and characterization of Ca2+-independent lipase from Bacillus pumilus B26. Biochimica et Biophys Acta (BBA)-Molecular Cell Biol Lipids 1583 (2):205–212.
4. **Reiss R, Ihssen J, Thöny-Meyer L**. 2011. Bacillus pumilus laccase: a heat stable enzyme with a wide substrate spectrum. BMC biotechnology 11 (1):1–11.
5. **Ahmadian G, Degrassi G, Venturi V, Zeigler D, Soudi M, Zanguinejad P**. 2007. Bacillus pumilus SG2 isolated from saline conditions produces and secretes two chitinases. J applied microbiology 103 (4):1081–1089.
6. **Reddy MR, Reddy KS, Chouhan YR, Bee H, Reddy G**. 2017. Effective feather degradation and keratinase production by Bacillus pumilus GRK for its application as bio-detergent additive. Bioresour technology 243:254–263.
7. **Wu Q, Li C, Li C, Chen H, Shuliang L**. 2010. Purification and characterization of a novel collagenase from Bacillus pumilus Col-J. Appl biochemistry biotechnology 160 (1):129–139.
8. **Baweja M, Tiwari R, Singh PK, Nain L, Shukla P**. 2016. An alkaline protease from Bacillus pumilus MP 27: functional analysis of its binding model toward its applications as detergent additive. Front microbiology 7:1195.
9. **Wang H, Liu D, Liu Y, Cheng C, Ma Q, Huang Q, Zhang Y**. 2007. Screening and mutagenesis of a novel Bacillus pumilus strain producing alkaline protease for dehairing. Lett applied microbiology 44 (1):1–6.
10. **Küppers T, Steffen V, Hellmuth H, O'Connell T, Bongaerts J, Maurer KH, Wiechert W**. 2014. Developing a new production host from a blueprint: Bacillus pumilus as an industrial enzyme producer. Microb cell factories 13 (1):1–11.
11. **De-Bashan LE, Hernandez JP, Bashan Y, Maier RM**. 2010. Bacillus pumilus ES4: candidate plant growth-promoting bacterium to enhance establishment of plants in mine tailings. Environ Exp Bot 69 (3):343–352.
12. **Hernandez JP, de Bashan LE, Rodriguez DJ, Rodriguez Y, Bashan Y**. 2009. Growth promotion of the freshwater microalga Chlorella vulgaris by the nitrogen-fixing, plant growth-promoting bacterium Bacillus pumilus from arid zone soils. european journal soil biology 45 (1):88–93.
13. **Gao XY, Liu Y, Miao LL, Li EW, Hou TT, Liu ZP**. 2017. Mechanism of anti-Vibrio activity of marine probiotic strain Bacillus pumilus H2, and characterization of the active substance. AMB Express 7 (1):1–10.
14. **Thy HTT, Tri NN, Quy OM, Fotedar R, Kannika K, Unajak S, Areechon N**. 2017. Effects of the dietary supplementation of mixed probiotic spores of Bacillus amyloliquefaciens 54A, and Bacillus pumilus 47B on growth, innate immunity and stress responses of striped catfish (Pangasianodon hypophthalmus). Fish & shellfish immunology 60:391–399.
15. **Khodzhaeva V, Makeeva A, Ulyanova V, Zelenikhin P, Evtugyn V, Hardt M, Rozhina E, Lvov Y, Fakhrullin R, Ilinskaya O**. 2017. Binase immobilized on halloysite nanotubes exerts enhanced cytotoxicity toward human colon adenocarcinoma cells. Front pharmacology 8:631.
16. **De Coster W, D'Hert S, Schultz DT, Cruts M, Van Broeckhoven C**. 2018. NanoPack: visualizing and processing long-read sequencing data. Bioinformatics 34 (15):2666–2669.
17. **Chen S, Zhou Y, Chen Y, Gu J**. 2018. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics 34 (17):i884–i890.
18. **Wick RR, Judd LM, Gorrie CL, Holt KE**. 2017. Unicycler: resolving bacterial genome assemblies from short and long sequencing reads. PLoS computational biology 13 (6):e1005595.
19. **Tatusova T, DiCuccio M, Badretdin A, Chetvernin V, Nawrocki EP, Zaslavsky L, Lomsadze A, Pruitt KD, Borodovsky M, Ostell J**. 2016.

ms Submission Template mSystems Submission Template mSystems Submission Template mSystems Submission Template mSystems Submission Template mSystems Submission Te

NCBI prokaryotic genome annotation pipeline. Nucleic acids research 44 (14):6614–6624.

20. **Pritchard L, Glover RH, Humphris S, Elphinstone JG, Toth IK**. 2016. Genomics and taxonomy in diagnostics for food security: soft-rotting enterobacterial plant pathogens. Anal Methods 8 (1):12–24.
21. **Blin K, Shaw S, Kloosterman AM, Charlop-Powers Z, van Wezel GP, Medema MH, Tilmann W**. 2021. antiSMASH 6.0. Nucleic Acids Res .
22. **Akhter S, Aziz RK, Edwards RA**. 2012. PhiSpy: a novel algorithm for finding prophages in bacterial genomes that combines similarity-and composition-based strategies. Nucleic acids research 40 (16):e126–e126.
23. **Xie Z, Tang H**. 2017. ISEScan: automated identification of insertion sequence elements in prokaryotic genomes. Bioinformatics 33 (21):3340–3347.
24. **Couvin D, Bernheim A, Toffano-Nioche C, Touchon M, Michalik J, Néron B, Rocha EP, Vergnaud G, Gautheret D, Pourcel C**. 2018. CRISPRCasFinder, an update of CRISRFinder, includes a portable version, enhanced performance and integrates search for Cas proteins. Nucleic acids research 46 (W1):W246–W251.

6.1 Additional work

The *Bacillus* genus is full of species with biotechnological application potential beyond those within the *B. subtilis* clade. As strains from *B. subtilis* and *B. licheniformis* dominate the field of industrial enzyme production, *B. thuringiensis* undeniably represents the bacterial champion in the fight against agricultural pests. *B. thuringiensis* takes protein production to the next level, by producing high amounts of pesticidal proteins efficiently packed as parasporal crystals. These proteins are widely implemented in crop protection strategies as environmentally friendly alternatives. The first publication of this section presents IDOPS, a tool developed to aid in the identification and comparative genomics study of pesticidal proteins. The second entry corresponds to a manuscript in preparation, which reports pesticidal sequences in an uncommon chromosomal location and associated with a prophage region.

frontiers in Microbiology

ORIGINAL RESEARCH
published: 28 June 2021
doi: 10.3389/fmicb.2021.664476

IDOPS, a Profile HMM-Based Tool to Detect Pesticidal Sequences and Compare Their Genetic Context

*Stefani Díaz-Valerio, Anat Lev Hacohen, Raphael Schöppe and Heiko Liesegang**

Genomic and Applied Microbiology & Göttingen Genomics Laboratory, Institute of Microbiology and Genetics, Georg-August University of Göttingen, Göttingen, Germany

Biopesticide-based crop protection is constantly challenged by insect resistance. Thus, expansion of available biopesticides is crucial for sustainable agriculture. Although *Bacillus thuringiensis* is the major agent for pesticide bioprotection, the number of bacteria species synthesizing proteins with biopesticidal potential is much higher. The Bacterial Pesticidal Protein Resource Center (BPPRC) offers a database of sequences for the control of insect pests, grouped in structural classes. Here we present IDOPS, a tool that detects novel biopesticidal sequences and analyzes them within their genetic environment. The backbone of the IDOPS detection unit is a curated collection of high-quality hidden Markov models that is in accordance with the BPPRC nomenclature. IDOPS was positively benchmarked with BtToxin_Digger and Cry_Processor. In addition, a scan of the UniProtKB database using the IDOPS models returned an abundance of new pesticidal protein candidates distributed across all of the structural groups. Gene expression depends on the genomic environment, therefore, IDOPS provides a comparative genomics module to investigate the genetic regions surrounding pesticidal genes. This feature enables the investigation of accessory elements and evolutionary traits relevant for optimal toxin expression and functional diversification. IDOPS contributes and expands our current arsenal of pesticidal proteins used for crop protection.

Keywords: biopesticide, hidden markov model, insecticidal protein, toxin identification, pesticidal, genetic context, IDOPS, comparative genomics

OPEN ACCESS

Edited by:
Tereza Castellane,
Sáo Paulo State University, Brazil

Reviewed by:
Ricardo Polanczyk,
Sáo Paulo State University, Brazil
Jie Zhang,
Institute of Plant Protection, Chinese Academy of Agricultural Sciences, China

****Correspondence:***
Heiko Liesegang
hlieseg@gwdg.de

Specialty section:
This article was submitted to Evolutionary and Genomic Microbiology, a section of the journal Frontiers in Microbiology

Received: *05 February 2021*
Accepted: *13 May 2021*
Published: *28 June 2021*

Citation:
Díaz-Valerio S, Lev Hacohen A, Schöppe R and Liesegang H (2021) IDOPS, a Profile HMM-Based Tool to Detect Pesticidal Sequences and Compare Their Genetic Context. Front. Microbiol. 12:664476. doi: 10.3389/fmicb.2021.664476

1. INTRODUCTION

Agricultural management strategies regard biopesticides as an environmentally friendly alternative to the chemical formulations used to suppress invertebrate pests (Mnif and Ghribi, 2015; Kachhawa, 2017). Plant-associated, soil, and entomopathogenic bacteria are a natural source of agents with pesticidal potential. Among those, *Bacillus thuringiensis* (Bt) strains and their derived crop protection products are safe for humans, highly specific to the targeted pests, and affordable to manufacture in bulk, making Bt the most successful biopesticide implemented worldwide (George and Crickmore, 2012; Jouzani et al., 2017). Nevertheless, nature fights back and target insects develop resistance mechanisms against Bt, which creates a constant need for novel and improved pest control agents (Vílchez, 2020).

Recently, it became clear that there is a wider variety of pesticidal proteins, synthesized by other bacteria that also interact with insects as part of their lifestyle (Castagnola and Stock, 2014; Ruiu, 2018). These have the potential to replace, supplement, and expand current options for biopest control (Waterfield et al., 2001). Bacteria like *Dickeya* spp., (Loth et al., 2015) *Yersinia* spp., and *Photorhabdus* spp. (Heermann and Fuchs, 2008) are now known sources of such proteins. The varied pesticidal proteins were re-classified in 16 structural groups by the Bacterial Pesticidal Protein Resource Center (BPPRC) (Crickmore et al., 2020a,b).

Biological databases are constantly enriched due to the proliferation of sequencing projects and advancements in sequencing technologies, thus they are a promising reservoir of sequences with uncharacterized pesticidal potential. However, screening such vast amounts of data requires sophisticated computational approaches in order to gain insight and take advantage of less explored resources. The use of profile hidden Markov models (HMMs) to analyze biological data has proven to be robust and sensitive (Eddy, 1998). A profile HMM condenses the information of a multiple alignment of homologous sequences, and therefore, has a higher discriminative power than pairwise similarity-based search tools like BLAST (Söding, 2005). Profile HMMs are broadly applied for protein family assignment, domain analysis, and detection of remote homologies in resources such as InterProScan (Jones et al., 2014; Blum et al., 2020), PFAM (Sonnhammer et al., 1998; Finn et al., 2007), and TIGRFAM (Haft et al., 2012). Examples of dedicated collections of profile HMMs are RVDB-prot (Bigot et al., 2019), a database for detection of viral proteins, and TASmania, a tool for the discovery of toxin-antitoxin systems in bacterial genomes (Akarsu et al., 2019).

Consequently, previous efforts to implement profile HMMs for detection of pesticidal proteins are not surprising. Cry_Processor (Shikov et al., 2020) is a tool for identification of 3 domain Cry sequences based on 4 profile HMMs, one for each domain and one full-length protein, making single domain delimitation possible. BtToxin_Digger (Liu et al., 2020), the successor of BtToxin_scanner, relies on a combination of HMMs, BLAST, and support vector machine (SVM) for prediction of not only 3 domain toxins but also members of the other structural groups.

The focus of the existing toxin prediction tools is to recognize and classify pesticidal protein sequences. Currently, none of them take into consideration the genetic environment of the genes coding for pesticidal proteins. Surrounding elements often include chaperones, crystallization domains, mobile elements, transporters, prophages, and virulence factors (Koni and Ellar, 1993; Shao et al., 2001; Elleuch et al., 2016; Adalat et al., 2017; Fayad et al., 2020; Lechuga et al., 2020). Moreover, the arrangement and distribution of such elements across genomes reveal crucial details about toxin functionality, host adaptation, diversification, and evolution of biopesticides (Khasdan et al., 2007; Peng et al., 2015; Ruffner et al., 2015; Fiedoruk et al., 2017; Zheng et al., 2017; Fayad et al., 2020; Wang et al., 2020). Hence, a more exhaustive approach would not only detect pesticidal sequences but also enable comparative genomics analysis of the candidate toxin in order to characterize the complete expression unit.

The goal of this study was to develop such tool. Our efforts originated IDOPS (Identification of Pesticidal Sequences), a software based on an extensive collection of high-quality profile HMMs. IDOPS aims to provide (i) a detection unit to aid in the finding of pesticidal proteins, especially novel variants within recently expanding groups, (ii) a basic classification system in accordance with the BPPRC nomenclature system, and (iii) a comparative genomics module to investigate toxin genes and their complete expression unit within their genomic environment.

2. MATERIALS AND METHODS

2.1. Building Profile HMMs

In order to better represent the great diversity of pesticidal proteins, we created a collection of known and putative pesticidal sequences. Our initial collection combined data from the previous *Bacillus thuringiensis* Toxin Nomenclature website (Crickmore et al., 1998), and matches for various pesticidal sequences found at UniProtKB-2020_06 (UniProt Consortium, 2019). Since UniProt-TrEMBL includes fragmented and repeated entries, redundancy was removed by clustering sequences of 100% identity with CD-HIT v.4.8.1 (Li and Godzik, 2006); then the longest member of each cluster was preserved. The representative sequences were used to create an all vs. all matrix with BLASTp v.2.9.0 (Camacho et al., 2009). This matrix was the input for clustering with the Markov Clustering Algorithm implemented by MCL (Enright et al., 2002). Resulting groups containing at least five members were aligned by ClustalO 1.2.4 (Sievers and Higgins, 2018), and the alignment was passed to hmmbuild-HMMER v.3.3 (Eddy, 1998) to create profile hidden Markov models (HMMs). Furthermore, an individual profile HMM was created to represent the C-terminal region of Cry toxins longer than 1,000 amino acids. For such purpose, the toxin core of the Cry sequences was removed from the alignment before the hmmbuild step. The preliminary profile HMM database accurately described pesticidal proteins of *B. thuringiensis* and was useful for detection of novel toxins. Nevertheless, at the time of this study, the BPPRC released the updated classification of pesticidal proteins and their source organisms. Therefore, the initial profile model collection undergone further examination and additional models were created to represent sequences from non Bt organisms.

2.2. Model Refinement and Validation

Several rounds of manual refinement and optimization were necessary to ensure high sensitivity and specificity of the models. This step included evaluation of the sequences used for each profile HMM, their phylogenetic relationship, domain signature predicted by the InterPro consortium (Blum et al., 2020), removal of biases produced by over-represented sequences and performance when databases were scanned using hmmsearch-HMMER v.3.3 (Eddy, 1998).

Protein sequences from the BPPRC database were considered as true positives. For a control dataset of true negatives, a collection of bacterial pore-forming toxins and related proteins from other bacteria was used (Gonzalez et al., 2008; **Supplementary Table 1**). The overall performance of the models was evaluated by scanning the UniProtKB databases. Furthermore, the distribution of the sequences matched by each model was analyzed and served to determine a trusted cutoff value above which true protein class members are found.

Criterion to define a good model are:

- Identification of all the true positive members of the protein class (or subclass) described by the model within the high score range.
- Additional matches within the high score range can be consistently assigned to the protein group by evaluation of domain signatures, source organism, and quality of alignment with true members.
- Distantly related pore-forming toxins from the true negative control dataset are not found at all or found below the trusted cutoff value.
- The distribution of the proteins matched by the model when searching UniProtKB databases indicates a clear separation between true members and other protein matches.

2.3. Pipeline Implementation

This collection of profile HMMs can be used effectively to search for matches within query sequences, genomes, and whole databases. It was implemented in a pipeline named Identification of Pesticidal Sequences (IDOPS). The software applied in IDOPS and the corresponding versions are presented in **Table 1**. To ease distribution and installation, IDOPS is available as a conda package (https://anaconda.org/GAMB-GO/idops) and its source code is found at Github (https://github.com/GAMB-GO/IDOPS), under a GPLv3 license.

Valid input formats for IDOPS are fasta (for single or multiple proteins) and genbank (for complete or draft genomes). Candidate pesticidal sequences are identified by hmmscan-HMMER v.3.3 (Eddy, 1998) and evaluated against the trusted cutoffs of the refined profile models. To facilitate the assessment of the found candidates, IDOPS internally generates profile alignments and reports phylogenetic trees for each match with the 10 nearest sequences of the corresponding protein groups.

Furthermore, in order to characterize the genomic context of identified pesticidal proteins and produce relevant insights regarding evolution and functionality, IDOPS offers an additional feature when genomic data is provided in genbank format. First, it retrieves the 5000 bp upstream and downstream regions of each match. Secondly, it does a Prokka annotation (Seemann, 2014) of such sequences. Finally, it creates an EasyFig (Sullivan et al., 2011) comparison that depicts BLAST identities between conserved regions.

TABLE 1 | Software dependencies and versions implemented in IDOPS.

Software	Version
Python	3.7.6
HMMer (hmmbuild, hmmsearch, hmmscan, hmmalign)	3.3
Biopython	1.76
Easyfig	2.2.5
Clustal Omega	1.2.4
BLAST	2.9.0
Prokka	1.14.6

2.4. Comparative Analysis of Genetic Environments

We used IDOPS to analyze eight *B. thuringiensis* plasmids carrying *cry1* genes (detailed strains and accession ID list in **Supplementary Table 2**). It was previously reported that *cry1* can occur alone or as part of an insecticidal pathogenicity island (Fiedoruk et al., 2017). We used this observation to demonstrate the value of IDOPS to facilitate the discovery of different genetic arrangements.

2.5. Benchmarking

The performance of IDOPS was compared with that of current tools for pesticidal protein detection: Cry_Processor (Shikov et al., 2020) and BtToxin_digger (Liu et al., 2020). Cry_Processor was developed to identify 3 domain Cry sequences. Therefore, the tool was examined taking in consideration only this group of proteins. It provides two search modes, Find Domains (FD) that is based upon a hmmsearch against generalized HMM models, and Domains Only (DO) which searches directly for each of the domains without a filtering step. The two modes were applied. Both tools were tested with sequences from the BPPRC as positive dataset and the true negative collection of distantly related pore-forming toxins.

3. RESULTS

3.1. A Collection of High-Quality Profile Hidden Markov Models

We developed highly specific profile HMMs to accurately represent each of the 16 structural groups defined by the Bacterial Pesticidal Protein Resource Center. The iterative refinement and manual curation process led to the creation of subgroups within some of the protein classes; particularly for highly populated and diverse groups, like Cry and Cyt. Our final collection consists of 31 profile hidden Markov models (detailed list in **Supplementary Table 3**).

The Xpp category for unclassified homology groups contains some proteins that could not be modeled based on multiple sequence alignments due to insufficient available sequences at the databases, namely Xpp37, Xpp76, and Xpp77. For those proteins single sequence models were created and incorporated in IDOPS. Furthermore, an individual profile HMM was dedicated to the C-terminal region of the 3 domain pesticidal proteins.

Every profile HMM satisfies the criteria for a good model. The final subsets of sequences selected to construct the models are enough to represent the whole diversity of each group while retaining specificity. Consequently, there is no significant overlap between the matches identified by each profile HMM, as shown at the **Supplementary Table 4**. None of the pore-forming proteins from the true negative dataset was identified by IDOPS models.

The scanning of UniProtKB databases was beneficial to elucidate the selective power of each model and to determine the gathering cutoff values. A density estimate of the hit distribution against UniProt-TrEMBL and UniProt-SwissProt depicts how

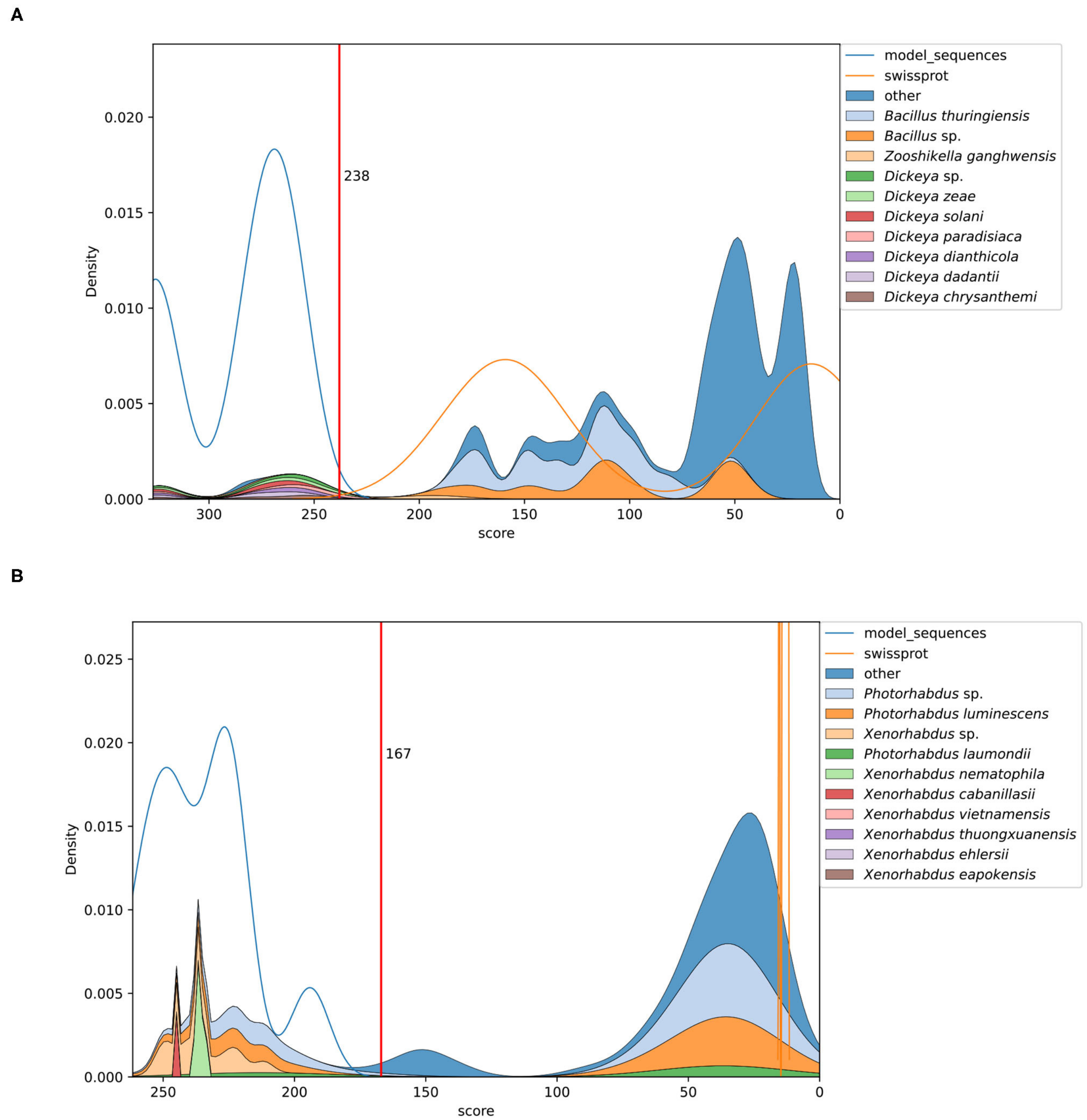


FIGURE 1 | Kernel density plots depicting distribution of model matches in UniProtKb. **(A)** The profile HMM identifies Cyt 4/5/6/7 proteins from *Dickeya* spp. within the high score range. The red line marks the curated gathering cutoff. Matches scoring below 238 correspond to related Cyt proteins from other source organisms. **(B)** A model targeting Pra sequences produced by *Photorhabdus* spp. and *Xenorhabdus* spp. Some *Yersinia* spp. proteins score directly next to the gathering cutoff.

accurately our models discriminate true members of each protein class from the whole database. In **Figure 1**, two examples are shown.

Figure 1A depicts the sequences identified by a model targeting Cyt 4/5/6/7 from *Dickeya* spp. Only proteins from *Dickeya* spp. score above the gathering cutoff and those below it correspond to related Cyt proteins produced by other organisms like *Bacillus thuringiensis*. The example in **Figure 1B** shows the matches of a model built to identify Pra proteins. Here, the high scoring sequences are clearly separated from most of the non-relevant hits. Interestingly, some entries, annotated as "uncharacterized proteins," score

TABLE 2 | Protein sequences identified by IDOPS profile HMMs in UniProtKb-TrEMBL compared with the amount of sequences at the Bacterial Pesticidal Protein Resource Center database.

	Number of protein Sequences above gathering cutoff	
Protein class	**BPPRC database**	**UniProtKB-TrEMBL**
app	10	121
cry	720	1,123
cyt	40	87
gpp	11	6
mcf	5	82
mpf	5	14
mpp	40	130
mtx	1	3
pra	3	59
prb	3	46
spp	2	482
tpp	30	52
vip	108	120
vpa	20	40
vpb	20	79
xpp	14	16

just below the gathering threshold. These are produced by members of the *Yersinia* genus. A similar distribution is observed when analyzing the matches of the toxin partner component, represented by the Prb HMM (Plots for each model are found as **Supplementary Figures 1–31**).

In addition, it is of notice the abundance of potential new pesticidal proteins identified by our approach. When the UniProtKb-TrEMBL database was scanned, a total of 2,460 protein sequences were found within the trusted score range of IDOPS profile HMMs, many of them annotated as uncharacterized or hypothetical proteins (**Table 2**).

Notably, models representing groups with few members in the BPPRC collection; like App and Spp, with 10 and 2 entries, respectively, detected plenty of potentially novel pesticidal sequences; 121 for App and 482 in the case of Spp. In other cases, no new entry is matched by our profile HMMs, this is true for Xpp76 and Xpp77 proteins, members of the Xpp group.

3.2. Comparative Analysis of Genetic Environments

In order to demonstrate the potential of IDOPS to investigate the genetic context of pesticidal proteins, we tested our tool against known reported *cry1* cassettes in plasmids of *B. thuringiensis* (**Figure 2**).

Using the sequences of eight Bt plasmids (retrieved in genbank format), IDOPS automatically generated a color coded and uniformly annotated alignment of genome regions encoding the *cry1* cassettes. In the red box, the figure shows the *cry1* cassette components (Cry1 -N-acetylmuramoyl-L-alanine amidase - K(+)/H(+) antiporter) and their genetic environment. IDOPS's output already allows the identification and grouping of the *cry1* cassette variants according to their particular elements, such as transposases from different families. Moreover, the figure is in accordance with the manually generated figure created by Fiedoruk et al. (2017), showing how IDOPS identifies, aligns, and displays the genomic surrounding of a toxin of interest.

3.3. Benchmarking

To evaluate the specificity and the sensitivity in comparison with currently implemented toxin identification tools, BtToxin_Digger and Cry_Processor, we applied benchmark tests (**Table 3**). The proteins from the BPPRC database were used as true positives and distantly related pore forming toxins sequences as a true negative dataset.

All three programs performed convincingly throughout the specificity test. None of the sequences from the true negative dataset were missannotated as a pesticidal protein by either Cry_Processor, BtToxin_Digger or IDOPS. The generalized profile HMMs of Cry_Processor returned some single domain matches. Nevertheless, those did not meet the criteria to be reported as pesticidal proteins.

Concerning the sensitivity Cry_Processor showed some discrepancy between the FD (Full Domain) and DO (Domain only) modes, as 696 and 715 out of 720 sequences were identified as 3 domain toxins, respectively (**Table 3**). In a similar way, BtToxin_Digger successfully recognized 994 of the 1,033 input sequences from the BPPRC. It failed to detect any of Spp and Vpb sequences, while identification of Cry, Vip, Mpp, Pra, and Tpp groups was incomplete (**Table 3**). IDOPS had the highest retrieval rate regarding this dataset. It recognized all but three Cry sequences above its gathering cutoff values. In the case of the missed toxins, Cry1Ca10, Cry3Bb3, and Cry11Aa2, a closer look revealed that these three sequences represent truncated toxins. The proteins sequences have been, nevertheless, recognized as hits that scored below trusted cut off. Exclusively IDOPS recovered all complete protein sequences from the tested true positive data set.

4. DISCUSSION

Here we present IDOPS, a tool to detect bacterial pesticidal protein sequences and compare their genetic environment. The power of IDOPS comes from a collection of high-quality profile hidden Markov models; each one carefully designed to represent a structural group as defined by the Bacterial Pesticidal Protein Resource Center (BPPRC) (Crickmore et al., 2020a,b). To this date, the tool comprises the most exhaustive and complete collection of models describing pesticidal proteins.

We compared IDOPS with other tools implementing profile HMMs, Cry_Processor (Shikov et al., 2020) and BtToxin_Digger (Liu et al., 2020). Neither of those recognized all of the complete sequences from the positive dataset, making the search against genomes or full databases potentially incomplete. Moreover, Cry_Processor's database is not up to date with the current BPPRC nomenclature. On the other hand, BtToxin_Digger has a greedy approach behind its profile HMMs. While this could work well for closely related and not so diverse structural groups, it is not the best option when dealing with varying sequences,

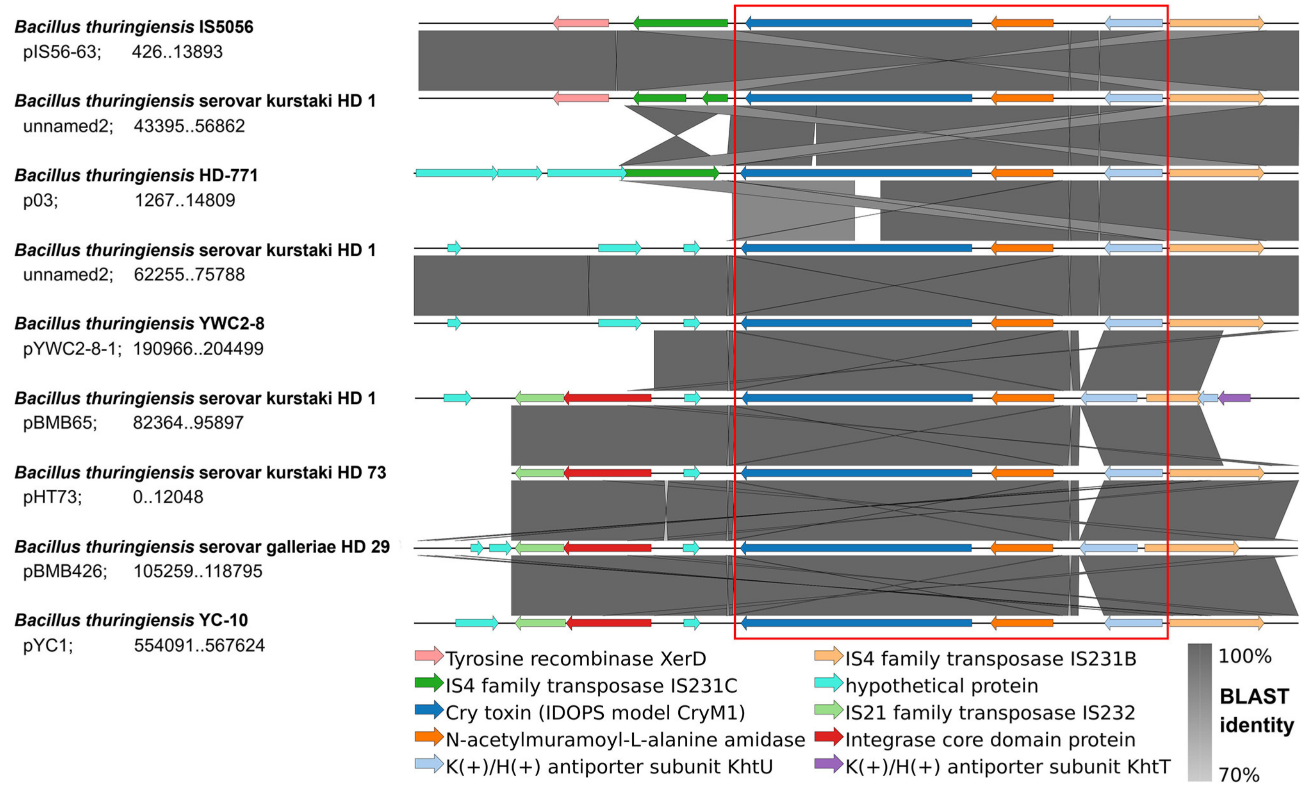

Bacillus thuringiensis IS5056
pIS56-63; 426..13893
Bacillus thuringiensis serovar kurstaki HD 1
unnamed2; 43395..56862
Bacillus thuringiensis HD-771
p03; 1267..14809
Bacillus thuringiensis serovar kurstaki HD 1
unnamed2; 62255..75788
Bacillus thuringiensis YWC2-8
pYWC2-8-1; 190966..204499
Bacillus thuringiensis serovar kurstaki HD 1
pBMB65; 82364..95897
Bacillus thuringiensis serovar kurstaki HD 73
pHT73; 0..12048
Bacillus thuringiensis serovar galleriae HD 29
pBMB426; 105259..118795
Bacillus thuringiensis YC-10
pYC1; 554091..567624
Tyrosine recombinase XerD
IS4 family transposase IS231C
Cry toxin (IDOPS model CryM1)
N-acetylmuramoyl-L-alanine amidase
K(+)/H(+) antiporter subunit KhtU
IS4 family transposase IS231B
hypothetical protein
IS21 family transposase IS232
Integrase core domain protein
K(+)/H(+) antiporter subunit KhtT
100%
BLAST identity
70%

FIGURE 2 | IDOPS genetic environment comparison. IDOPS retrieved, annotated, and aligned the genetic region surrounding *cry1* genes reported as a cassette within Bt plasmids by Fiedoruk et al. (2017). The visualization is generated by EasyFig as implemented internally in IDOPS. The reported *cry1* cassette is marked in red. Strain names, plasmid, and coordinates appear next to each row.

such as members of the Cry group or even proteins of the Xpp class, which lack in shared homology between them. IDOPS overcomes such limitations with its comprehensive collection of profile HMMs. Additionally, it provides a unique genetic context comparison feature.

Gathering cutoff values for IDOPS' models are optimized to recognize full-length proteins, thus shorter or incomplete sequences will score below it and won't be reported. In our tests, IDOPS recognized 717 from the 720 Cry sequences at the BPPRC database. A detailed look revealed that none of the three missed sequences are full-length 3 domain proteins. Cry3Bb3 presents only the InterPro signatures of the central (IPR001178) and the C-terminal domains (IPR005638). Moreover, Cry1Ca10 and Cry11Aa2 are both annotated as partial proteins with sequence lengths of 181aa and 78aa, respectively. Cry1Cb3 is a particular case, since it is not a full-length toxin but rather just the C-terminal portion of a long Cry protein, with the C-terminal (IPR005638) and domain V (IPR041587) regions. Since we developed a dedicated model for the C-terminal region of a 3 domain protein, and the model recognized the sequence as such, it was reported, but under this considerations. However, in case a researcher using IDOPS needs to retrieve the low-scoring proteins, we set up an option to disable the gathering cutoff, so even incomplete matches will be reported for further manual evaluation.

Bias within biological databases affects the creation and refinement of profile HMMs. Cry is one of the most studied pesticidal protein groups, as reflected by the amount of available sequences and their diversity, which are valuable to build rich and diverse profile HMMs. Nevertheless, there is a composition bias within the available Cry sequences. The BPPRC database contains 720 Cry entries and 276 of those correspond to Cry1 sequences. Such skewed composition may be carried over to further studies and databases. To ensure that the profile HMMs do not suffer from this bias, several rounds of model training and refinement were done to find the adequate sequences to represent each pesticidal class. Conversely, other groups, such as Mcf, Mtx, and Spp have significantly fewer representatives, sometimes single entries; this in turn makes model building a challenging task.

IDOPS' models take into consideration common properties of distinct subgroups within each pesticidal protein group. For instance, Cyt proteins that are synthesized by the *Dickeya* spp. cluster separately from the well resolved Cyt proteins of the *Bacillus* clade. They have a shorter N-terminal region and lack hemolytic activity when compared with Cyts from the *Bacillus* spp. (Soberón et al., 2013; Loth et al., 2015). Consequently, it becomes reasonable to have a distinct model to better represent this subgroup. In a similar way, the Pra and Prb proteins of *Vibrio* spp. were modeled separately from their counterparts found in *Xenorhabdus* spp. and *Photorhabdus* spp.

TABLE 3 | Benchmark of tools identifying pesticidal sequences from the Bacterial Pesticidal Protein Resource Center (BPPRC).

Structural group	Sequences at the BPPRC	IDOPS	BtToxin _digger	Cry_processor	
				full_domain	domain_only
Cry	720	716	706	696	715
App	10	10	10	N\A	
Cyt	40	40	40	N\A	
Gpp	11	11	11	N\A	
Mcf	5	5	5	N\A	
Mpf	5	5	5	N\A	
Mpp	40	40	39	N\A	
Mtx	1	1	1	N\A	
Pra	3	3	2	N\A	
Prb	3	3	3	N\A	
Spp	2	2	0	N\A	
Tpp	30	30	30	N\A	
Vip	108	108	107	N\A	
Vpa	20	20	20	N\A	
Vpb	20	20	0	N\A	
Xpp	14	14	14	N\A	

The current BPPRC database contains three entries for the Pra category, which corresponds to "*Photorhabdus* Insect-Related toxin A component" produced by *Photorhabdus luminescens* subsp. *luminescens*, *Xenorhabdus nematophila*, and *Vibrio parahaemolyticus* M0605 (Crickmore et al., 2020a). Nevertheless, it is intriguing to find sequences from *Yersinia* spp. scoring very close to the BPPRC proteins. *Yersinia* spp. shares insecticidal potential with *P. luminescens*, as homologous proteins with similar genetic arrangements have been found, perhaps as product of horizontal gene transfer (Heermann and Fuchs, 2008; Ahantarig et al., 2009; Castagnola and Stock, 2014; Hurst et al., 2016). Therefore, the sequences encoded by *Yersinia* spp. and matched by the Pra model might represent related variants of the Pra toxin. Supporting this idea, the match distribution of the Prb model shows *Yersinia* spp. proteins in a similar position, meaning this bacteria might potentially produce both the PirA and the PirB toxin components.

Besides source organisms, structural variations within each pesticidal protein group were contemplated for IDOPS' models. Cry2, Cry11, and Cry18, despite being part of the 3 domain category, lack some of the conserved blocks described for this group (Schnepf et al., 1998; Palma et al., 2014). Accordingly, a distinct model was created for the proteins of this subgroup. In a similar way, another subgroup of Cry proteins present variant and alternate versions of such blocks (Schnepf et al., 1998; de Maagd et al., 2001); for example, Cry5, Cry12, and Cry21, thus they were grouped and modeled apart.

IDOPS provides a profile HMM dedicated to the C-terminal extension of the Cry proteins. The rationale for its creation is the evidence of such sequences encoded in proximity to the short variants of *cry* genes. These proteins have homology to the C-terminal region of the long Cry toxins (de Maagd et al., 2003). A role in crystal formation, packing, and stabilization has been reported for the C-terminal extension (Naimov et al., 2006; Peng et al., 2015). Moreover, chimeric toxins made by artificial recombination of N-terminal and C-terminal regions of Cry proteins have shown increased crystal stability and toxicity (Naimov et al., 2006; Zghal et al., 2016). Therefore, with the C-terminal model, IDOPS contributes to identify independent instances of this C-terminal extension. These instances may be useful to investigate crystallization properties and toxic activity of pesticidal proteins.

IDOPS' profile HMMss were meticulously tested against a) the sequence collection at BPPRC, b) a true negative dataset of pore-forming toxins, and c) the whole UniProtKB database. IDOPS recognized all complete sequences of the true positive dataset, none of the false positives, and 2,460 further sequences with pesticidal potential from the UniProtKB database. The abundance of sequences identified at UniProtKB exposes the unexplored potential of the less investigated pesticidal groups (**Table 1**). Having such dedicated models allows to infer some aspects regarding the candidate pesticidal protein identified by IDOPS. For example, whether it is: a Cry with the conserved blocks, a member of Xpp of a specific subtype, or to which kind of Cyt it belongs. Altogether, the final models are sensitive, specific for each structural group and constitute a promising aid in the search for novel pesticidal protein sequences.

Further developments of IDOPS will include expansion of the search toward other sequence collections, such as metagenomic data. By extended scanning, novel members of the less populated groups could be detected and used to improve the current single sequences models such as Xpp37, Xpp76, and Xpp77. Moreover, as other virulence factors have been shown to support the toxic activity of pesticidal proteins, especially in *B. thuringiensis* (George and Crickmore, 2012; Malovichko et al., 2019), it may be worth to target some of these features by dedicated profile HMMs. Another course of action for IDOPS will be the implementation of the tool as a website service to facilitate its access to the scientific community without the need of local installation.

IDOPS facilitates the comparison of the genetic environment of pesticidal sequences in a systematic and reliable way. The analysis of plasmids carrying *cry1* genes automatically generated an output consistent with Fiedoruk et al. (2017) results. Comparative genomics have proven relevant to understand genetic dynamics of pesticidal proteins. For example, Lechuga et al. (2020) proposed a plasmidial origin to a chromosomally-located Cry1Ba4 only after examination and comparison of the genetic context with that of plasmid-encoded toxins. In a further extensive analysis, we used IDOPS to detect a previously unreported chromosomal *cry* cassette. By comparative genomics we discovered *cry5*, *cry10*, and *cry13* variants in a rather highly conserved genetic environment. Moreover, we consistently found a *Siphoviridae*-like prophage region in the vicinity of the cassette (Lev Hacohen et al. unpublished data). IDOPS was of great help to detect such arrangement in several Bt strains, opening intriguing evolutionary questions.

We achieved a sophisticated and comprehensive tool that provides not only detection and structural classification of pesticidal proteins, but also a feature that identifies, retrieves, and aligns the genomic context of the pesticidal sequences. IDOPS was designed in accordance to the BPPRC nomenclature system. The benchmark confirmed it has the highest sensitivity available among other toxin search tools. This combination of a highly sensitive toxin detection engine with a solid genome comparison module is, to our knowledge, unique. All considered, we created IDOPS as a tool that addresses comparative genomics of pesticidal proteins to gain novel insights on biopesticides.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/**Supplementary Material**.

AUTHOR CONTRIBUTIONS

SD-V build the initial profile HMMs, performed research, analyzed data, and wrote the manuscript. ALH performed research, refined profile HMMs, and wrote the manuscript. RS performed research, coded the software tool, and wrote the manuscript. HL designed the study, performed research, evaluated results, and wrote the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

We acknowledge support by the Open Access Publication Funds of the Göttingen University.

ACKNOWLEDGMENTS

We thank Drs. Cynthia Maria Chibani and Bernd Wemheuer for intense scientific discussions. We thank Tarek Morsi for expert IT support with the G2L servers. Last but not least, we are grateful to Prof. Rolf Daniel for hosting and supporting us and giving us unlimited access to the G2L-IT infrastructure.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2021.664476/full#supplementary-material

REFERENCES

Adalat, R., Saleem, F., Crickmore, N., Naz, S., and Shakoori, A. R. (2017). *In vivo* crystallization of three-domain cry toxins. *Toxins* 9, 80. doi: 10.3390/toxins9030080

Ahantarig, A., Chantawat, N., Waterfield, N. R., Ffrench-Constant, R., and Kittayapong, P. (2009). Pirab toxin from *Photorhabdus asymbiotica* as a larvicide against dengue vectors. *Appl. Environ. Microbiol.* 75, 4627–4629. doi: 10.1128/AEM.00221-09

Akarsu, H., Bordes, P., Mansour, M., Bigot, D.-J., Genevaux, P., and Falquet, L. (2019). Tasmania: a bacterial toxin-antitoxin systems database. *PLoS Comput. Biol.* 15:e1006946. doi: 10.1371/journal.pcbi.1006946

Bigot, T., Temmam, S., Pérot, P., and Eloit, M. (2019). Rvdb-prot, a reference viral protein database and its hmm profiles. *F1000Research* 8, 530. doi: 10.12688/f1000research.18776.1

Blum, M., Chang, H.-Y., Chuguransky, S., Grego, T., Kandasaamy, S., Mitchell, A., et al. (2020). The interpro protein families and domains database: 20 years on. *Nucleic Acids Res.* 49, D344–D354. doi: 10.1093/nar/gkaa977

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., et al. (2009). Blast+: architecture and applications. *BMC Bioinformatics* 10:421. doi: 10.1186/1471-2105-10-421

Castagnola, A., and Stock, S. P. (2014). Common virulence factors and tissue targets of entomopathogenic bacteria for biological control of lepidopteran pests. *Insects* 5, 139–166. doi: 10.3390/insects5010139

Crickmore, N., Berry, C., Panneerselvam, S., Mishra, R., Connor, T. R., and Bonning, B. (2020a). *Bacterial Pesticidal Protein Resource Center*. Available online at: https://www.bpprc.org/

Crickmore, N., Berry, C., Panneerselvam, S., Mishra, R., Connor, T. R., and Bonning, B. C. (2020b). A structure-based nomenclature for *Bacillus thuringiensis* and other bacteria-derived pesticidal proteins. *J. Invertebr. Pathol.* 107438. doi: 10.1016/j.jip.2020.107438

Crickmore, N., Zeigler, D. R., Feitelson, J., Schnepf, E., Van Rie, J., Lereclus, D., et al. (1998). Revision of the nomenclature for the bacillus thuringiensis pesticidal crystal proteins. *Microbiol. Mol. Biol. Rev.* 62, 807–813.

de Maagd, R. A., Bravo, A., Berry, C., Crickmore, N., and Schnepf, H. E. (2003). Structure, diversity, and evolution of protein toxins from spore-forming entomopathogenic bacteria. *Ann. Rev. Genet.* 37, 409–433. doi: 10.1146/annurev.genet.37.110801.143042

de Maagd, R. A., Bravo, A., and Crickmore, N. (2001). How *Bacillus thuringiensis* has evolved specific toxins to colonize the insect world. *Trends in Genet.* 17, 193–199. doi: 10.1016/S0168-9525(01)02237-5

Eddy, S. R. (1998). Profile hidden markov models. *Bioinformatics* 14, 755–763. doi: 10.1093/bioinformatics/14.9.755

Elleuch, J., Jaoua, S., Ginibre, C., Chandre, F., Tounsi, S., and Zghal, R. Z. (2016). Toxin stability improvement and toxicity increase against dipteran and lepidopteran larvae of *Bacillus thuringiensis* crystal protein cry2aa. *Pest Manag. Sci.* 72, 2240–2246. doi: 10.1002/ps.4261

Enright, A. J., Van Dongen, S., and Ouzounis, C. A. (2002). An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.* 30, 1575–1584. doi: 10.1093/nar/30.7.1575

Fayad, N., Kambris, Z., El Chamy, L., Mahillon, J., and Awad, M. K. (2020). A novel anti-dipteran *bacillus thuringiensis* strain: Unusual cry toxin genes in a highly dynamic plasmid environment. *Appl. Environ. Microbiol.* 87:e02294-20. doi: 10.1128/AEM.02294-20

Fiedoruk, K., Daniluk, T., Mahillon, J., Leszczynska, K., and Swiecicka, I. (2017). Genetic environment of cry1 genes indicates their common origin. *Genome Biol. Evol.* 9, 2265–2275. doi: 10.1093/gbe/evx165

Finn, R. D., Tate, J., Mistry, J., Coggill, P. C., Sammut, S. J., Hotz, H.-R., et al. (2007). The pfam protein families database. *Nucleic Acids Res.* 36(Suppl. 1):D281–D288. doi: 10.1093/nar/gkm960

George, Z., and Crickmore, N. (2012). "*Bacillus thuringiensis* applications in agriculture," in *Bacillus thuringiensis Biotechnology* (Dordrecht: Springer), 19–39. doi: 10.1007/978-94-007-3021-2_2

Gonzalez, M., Bischofberger, M., Pernot, L., Van Der Goot, F., and Freche, B. (2008). Bacterial pore-forming toxins: the (w) hole story? *Cell. Mol. Life Sci.* 65, 493–507. doi: 10.1007/s00018-007-7434-y

Haft, D. H., Selengut, J. D., Richter, R. A., Harkins, D., Basu, M. K., and Beck, E. (2012). Tigrfams and genome properties in 2013. *Nucleic Acids Res.* 41, D387–D395. doi: 10.1093/nar/gks1234

Heermann, R., and Fuchs, T. M. (2008). Comparative analysis of the *Photorhabdus luminescens* and the *Yersinia enterocolitica* genomes: uncovering candidate genes involved in insect pathogenicity. *BMC Genomics* 9:40. doi: 10.1186/1471-2164-9-40

Hurst, M. R., Beattie, A., Altermann, E., Moraga, R. M., Harper, L. A., Calder, J., et al. (2016). The draft genome sequence of the *Yersinia entomophaga* entomopathogenic type strain mh96t. *Toxins* 8, 143. doi: 10.3390/toxins8050143

Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C., et al. (2014). Interproscan 5: genome-scale protein function classification. *Bioinformatics* 30, 1236–1240. doi: 10.1093/bioinformatics/btu031

Jouzani, G. S., Valijanian, E., and Sharafi, R. (2017). *Bacillus thuringiensis*: a successful insecticide with new environmental features and tidings. *Appl Microbiol. Biotechnol.* 101, 2691–2711. doi: 10.1007/s00253-017-8175-y

Kachhawa, D. (2017). Microorganisms as a biopesticides. *J. Entomol. Zool. Stud.* 5, 468–473.

Khasdan, V., Sapojnik, M., Zaritsky, A., Horowitz, A. R., Boussiba, S., Rippa, M., et al. (2007). Larvicidal activities against agricultural pests of transgenic *Escherichia coli* expressing combinations of four genes from *Bacillus thuringiensis*. *Arch. Microbiol.* 188, 643–653. doi: 10.1007/s00203-007-0285-y

Koni, P., and Ellar, D. (1993). Cloning and characterization of a novel *Bacillus thuringiensis* cytolytic delta-endotoxin. *J. Mol. Biol.* 229, 319–327. doi: 10.1006/jmbi.1993.1037

Lechuga, A., Lood, C., Salas, M., van Noort, V., Lavigne, R., and Redrejo-Rodríguez, M. (2020). Completed genomic sequence of *Bacillus thuringiensis* her1410 reveals a cry-containing chromosome, two megaplasmids, and an integrative plasmidial prophage. *G3* 10, 2927–2939. doi: 10.1534/g3.120.401361

Li, W., and Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22, 1658–1659. doi: 10.1093/bioinformatics/btl158

Liu, H., Zheng, J., Yu, Y., Ye, W., Peng, D., and Sun, M. (2020). Bttoxin_digger: a comprehensive and high-throughput pipeline for mining toxin protein genes from *Bacillus thuringiensis*. *bioRxiv*. doi: 10.1101/2020.05.26.114520

Loth, K., Costechareyre, D., Effantin, G., Rahbé Y., Condemine, G., Landon, C., and Da Silva, P. (2015). New cyt-like δ-endotoxins from *Dickeya dadantii*: structure and aphicidal activity. *Sci. Rep.* 5, 1–10. doi: 10.1038/srep08791

Malovichko, Y. V., Nizhnikov, A. A., and Antonets, K. S. (2019). Repertoire of the bacillus thuringiensis virulence factors unrelated to major classes of protein toxins and its role in specificity of host-pathogen interactions. *Toxins* 11, 347. doi: 10.3390/toxins11060347

Mnif, I., and Ghribi, D. (2015). Potential of bacterial derived biopesticides in pest management. *Crop Protect.* 77, 52–64. doi: 10.1016/j.cropro.2015.07.017

Naimov, S., Martens-Uzunova, E., Weemen-Hendriks, M., Dukiandjiev, S., Minkov, I., and de Maagd, R. A. (2006). Carboxy-terminal extension effects on crystal formation and insecticidal properties of colorado potato beetle-active bacillus thuringiensis δ-endotoxins. *Mol. Biotechnol.* 32, 185–196. doi: 10.1385/MB:32:3:185

Palma, L., Muñoz, D., Berry, C., Murillo, J., and Caballero, P. (2014). *Bacillus thuringiensis* toxins: An overview of their biocidal activity. *Toxins* 6, 3296–3325. doi: 10.3390/toxins6123296

Peng, D.-h., Pang, C.-y., Wu, H., Huang, Q., Zheng, J.-S., and Sun, M. (2015). The expression and crystallization of cry65aa require two c-termini, revealing a novel evolutionary strategy of *Bacillus thuringiensis* cry proteins. *Sci. Rep.* 5:8291. doi: 10.1038/srep08291

Ruffner, B., Péchy-Tarr, M., Höfte, M., Bloemberg, G., Grunder, J., Keel, C., and Maurhofer, M. (2015). Evolutionary patchwork of an insecticidal toxin shared between plant-associated pseudomonads and the insect pathogens *Photorhabdus* and *Xenorhabdus*. *BMC Genomics* 16:609. doi: 10.1186/s12864-015-1763-2

Ruiu, L. (2018). Microbial biopesticides in agroecosystems. *Agronomy* 8, 235. doi: 10.3390/agronomy8110235

Schnepf, E., Crickmore, N., Van Rie, J., Lereclus, D., Baum, J., Feitelson, J., et al. (1998). *Bacillus thuringiensis* and its pesticidal crystal proteins. *Microbiol. Mol. Biol. Rev.* 62, 775–806. doi: 10.1128/MMBR.62.3.775-806.1998

Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. *Bioinformatics* 30, 2068–2069. doi: 10.1093/bioinformatics/btu153

Shao, Z., Liu, Z., and Yu, Z. (2001). Effects of the 20-kilodalton helper protein on cry1ac production and spore formation in *Bacillus thuringiensis*. *Appl. Environ. Microbiol.* 67, 5362–5369. doi: 10.1128/AEM.67.12.5362-5369.2001

Shikov, A. E., Malovichko, Y. V., Skitchenko, R. K., Nizhnikov, A. A., and Antonets, K. S. (2020). No more tears: mining sequencing data for novel bt cry toxins with cryprocessor. *Toxins* 12, 204. doi: 10.3390/toxins12030204

Sievers, F., and Higgins, D. G. (2018). Clustal omega for making accurate alignments of many protein sequences. *Protein Sci.* 27, 135–145. doi: 10.1002/pro.3290

Soberón, M., López-Díaz, J. A., and Bravo, A. (2013). Cyt toxins produced by *Bacillus thuringiensis*: a protein fold conserved in several pathogenic microorganisms. *Peptides* 41:87–93. doi: 10.1016/j.peptides.2012.05.023

Söding, J. (2005). Protein homology detection by hmm–hmm comparison. *Bioinformatics* 21, 951–960. doi: 10.1093/bioinformatics/bti125

Sonnhammer, E. L., Eddy, S. R., Birney, E., Bateman, A., and Durbin, R. (1998). Pfam: multiple sequence alignments and hmm-profiles of protein domains. *Nucleic Acids Res.* 26, 320–322. doi: 10.1093/nar/26.1.320

Sullivan, M. J., Petty, N. K., and Beatson, S. A. (2011). Easyfig: a genome comparison visualizer. *Bioinformatics* 27, 1009–1010. doi: 10.1093/bioinformatics/btr039

UniProt Consortium (2019). Uniprot: a worldwide hub of protein knowledge. *Nucleic Acids Res.* 47, D506–D515. doi: 10.1093/nar/gky1049

Vílchez, S. (2020). Making 3d-cry toxin mutants: Much more than a tool of understanding toxins mechanism of action. *Toxins* 12, 600. doi: 10.3390/toxins12090600

Wang, Z., Wang, K., Bravo, A., Soberón, M., Cai, J., Shu, C., et al. (2020). Coexistence of cry9 with the vip3a gene in an identical plasmid of *Bacillus thuringiensis* indicates their synergistic insecticidal toxicity. *J. Agric. Food Chem.* 68, 14081–14090. doi: 10.1021/acs.jafc.0c05304

Waterfield, N. R., Bowen, D. J., Fetherston, J. D., Perry, R. D., et al. (2001). The tc genes of *Photorhabdus*: a growing family. *Trends Microbiol.* 9, 185–191. doi: 10.1016/S0966-842X(01)01978-3

Zghal, R. Z., Elleuch, J., Ali, M. B., Darriet, F., Rebaï, A., Chandre, F., et al. (2016). Towards novel cry toxins with enhanced toxicity/broader: a new chimeric cry4ba / cry1ac toxin. *Appl. Microbiol. Biotechnol.* 101, 113–122. doi: 10.1007/s00253-016-7766-3

Zheng, J., Gao, Q., Liu, L., Liu, H., Wang, Y., Peng, D., et al. (2017). Comparative genomics of *Bacillus thuringiensis* reveals a path to specialized exploitation of multiple invertebrate hosts. *MBio* 8:e00822-17. doi: 10.1128/mBio.00822-17

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2021 Díaz-Valerio, Lev Hacohen, Schöppe and Liesegang. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Article

Comparative Genomics of Chromosomally-Encoded Pesticidal Genes Reveals a Novel Prophage-Associated *cry* Cassette

Anat Lev Hacohen [1,†], Stefani Díaz Valerio [1,†], Jacqueline Hollensteiner [1], Raphael Schöppe [1] and Heiko Liesegang [1,‡,*]

[1] Genomic and Applied Microbiology & Göttingen Genomics Laboratory, Institute of Microbiology and Genetics, Georg-August University of Göttingen, Göttingen, Germany
* Correspondence: hlieseg@gwdg.de; Tel.: +49-551-39-9839
† These authors contributed equally to this work.
‡ Institute for Microbiology and Genetics, Grisebachstr. 8, D-37077 Göttingen, Germany

Abstract: *Bacillus thuringiensis* is known for the production of a large variety of toxic proteins, which are widely used as biopesticides in crop protection strategies. These pesticidal toxins are usually encoded on conjugative plasmids and exposed to various recombination events that can lead to improved host adaptation and increased fitness. Nevertheless, some studies have identified rare instances of chromosomally-encoded toxins. These unique cases raise evolutionary questions in regards to how and why were they established. In this comparative genomics study, chromosomally-encoded three-domain Cry toxins within genomes of *B. thuringiensis* were systematically investigated. The analysis was done using IDOPS, an in-house developed software designed to identify bacterial pesticidal proteins and compare their genetic environments. We identified (i) a novel genetic cassette consisting of two toxin-associated genes that envelope a coding sequence of a three-domain Cry toxin, (ii) evidence of recombination events resulting in various three-domain Cry toxins encoded by either one or two separated genes, and (iii) cassette-associated *Siphoviridae*-like prophage regions. These observations might hold the key for understanding the evolution of chromosomally-encoded Cry toxins.

Keywords: pesticidal toxins, crystal toxins, chromosomally-encoded toxins, comparative genomics, prophage-associated, genomic *cry* cassette

Key Contribution: This is, to our knowledge, the first comparative genomics study on chromosomally-encoded *cry* genes from *Bacillus thuringiensis*. Furthermore, the study highlight the proximity of a prophage region to a novel *cry* cassette, which may imply a prophage role in the transmission of toxin genes.

Citation: Lev Hacohen, A.; Díaz Valerio, S.; Hollensteiner, J.; Schöppe, R.; Liesegang, H. Comparative Genomics of Chromosomally-Encoded Pesticidal Genes Reveals a Novel Prophage-Associated *cry* Cassette. *Toxins* **2021**, *1*, 0. https://doi.org/

Received:
Accepted:
Published:

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2022 by the authors. Submitted to *Toxins* for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/ 4.0/).

1. Introduction

Bacillus thuringiensis (Bt), a member of the *Bacillus cereus* (Bc) *sensu lato* group [1], are a Gram-positive, spore-forming, ubiquitously distributed bacteria that inhabit soil, plants, and insect-related environments [2]. Toxins produced by Bt have high specificity towards numerous invertebrates such as insects, nematodes, and snails [3]. The Bacterial Pesticidal Protein Resource Center (BPPRC) recently released a structure-based nomenclature system for toxins synthesized by Bt and other insecticidal bacteria [4]. Bt produces toxins of the major structural groups Cry (three-domain crystal proteins), Cyt (cytolytic), Vip (vegetative insecticidal protein), and Mpp (ETX/Mtx2 family). Among those, the crystal forming three-domain Cry toxins represent the most prominent group of insecticidal proteins [5]. As a result of their safety to both humans and plants, and their straightforward bulk production, the use of these toxins as biopesticides has been ongoing for decades [6]. Bt-based products dominate the biopesticide market as a widely applied crop protection method as well as an agent in the battle against diseases-spreading mosquitoes [7]. Besides toxins, Bt strains also encode additional virulence factors, such as phospholipase C, proteases, and hemolysins, which support the toxicity against various target organisms [8].

The increasing availability of high-quality Bt genomes revealed a large number of toxin genes encoded by extrachromosomal elements. Previous studies have shown that many of these toxin-encoding plasmids are transferred between Bt strains by means of conjugation and undergo further divergence evolution processes through homologous recombination [3,9–11]. These, in turn, contribute to an improved host adaptation and increased fitness. *Bacillus thuringiensis* is known for the production of a large variety of toxic proteins, which are widely used as biopesticides in crop protection strategies. These pesticidal toxins are usually encoded on conjugative plasmids and are exposed to various recombination events that can lead to improved host adaptation and increased fitness. Recent studies reported rare instances of chromosomally-encoded Bt toxins [12,13], raising questions about diversification, functionality, and alternative ways of aquisition of the pesticidal genes. For example, three non-identical *cry13* genes were identified within the chromosome of Bt MYBT18246 by [13]. Moreover, each of these *cry13* gene was found in close proximity to a prophage region. Phages represent an abundant class of mobile genetic elements within the Bc *sensu lato* group and play a major role in the evolution and virulence of bacterial pathogens [9,14,15].

In order to further characterize unusual instances of chromosomally encoded Cry toxins, their genetic environment was investigated by comparative genomics approaches. The focus was to identify and describe chromosomal *cry* gene instances and to elucidate shared common elements surrounding them across Bt genomes. This study aimed to provide insight about the selective pressures and the molecular entities that resulted in the evolution of chromosomally-encoded cry genes. Considering that successful pesticidal Bt strains in naturally competitive environments were generated by such mechanisms, the insights provided by this investigation offers researchers alternatives to improve current biopesticidal strains.

2. Results

2.1. Comparative genomics of prophage-associated cry cassettes in B. thuringiensis

A novel genetic arrangement was discovered within complete genomes of *B. thuringiensis* by a comparative genomics approach. The pattern was detected for the three reported chromosomal *cry13* genes of Bt MYBT18246 [13] and also for the known chromosomal *cry5* gene in Bt YBT-1518 [12]. In all four cases, a set of three components co-occur. The first is a gene that encodes a protein of unknown function with a potential helix-turn-helix motif (*puf*). The second is a pesticidal gene encoding a three-domain Cry toxin. The third is an N-acetylmuramoyl-L-alanine amidase encoding-gene (*ami*). Bt MYBT18246, as opposed to Bt YBT-1518, presents a split variant of the *cry* gene, with a short toxin core coding sequence and a separate C-terminal coding sequence (*ter*). In all four cases, this genetic pattern is located upstream of a prophage region. This genomic arrangement was, therefore, defined as a prophage-associated *cry* cassette (**Figure 1**).

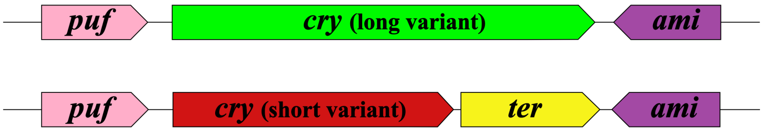


Figure 1. Schematic representation of the prophage-associated *cry* cassette. The cassette consists of a gene coding a protein of unknown function (*puf*), an amidase-encoding gene (*ami*), and a *cry* gene that can either be found as a long variant or as a split variant with a short toxin core coding sequence and a separate C-terminal coding sequence (*ter*).

A customized BLASTn [16] search masking the *cry* nucleotide sequence, was done in order to retrieve further *cry* gene variants surrounded by the same *puf* and *ami* genes observed in the prophage-associated *cry* cassettes of Bt MYBT18246 and Bt YBT-1518. The search returned eight additional cases of the complete prophage-associated *cry* cassette and five additional partial cassettes that appeared at the end of a contig and were, therefore, only suspected prophage-associated *cry* cassettes. **Table 1** lists the identified strains that contain the full prophage-associated *cry* cassette (**Table S2** contains the full list, including the partial prophage-associated *cry* cassettes). All bacterial strains presenting the reported cassette, except for one, were found within genomes of the *B. thuringiensis* species. The exceptional instance was detected in *B. cereus* VD154.

Table 1: *Bacillus* strains encoding the novel prophage-associated *cry* cassette.

Strain	Genome Status	Cassette Localization	Accession	Encoded Toxin	Toxic Against
Bc VD154_1 (+)	draft	contig 30	AHFG01000030.1	Cry5	nematodes
Bc VD154_2 (+)	draft	contig 37	AHFG01000037.1	Cry5	nematodes
Bc VD154_3 (+)	draft	contig 55	AHFG01000055.1	Cry65	not yet known
Bt AFS063736 (+)	draft	contig 59	NVCX01000059.1	Cry65	not yet known
Bt F14-1 (+)	draft	contig 26	JZKB01000011.1	Cry5	nematodes
Bt MYBT18246_1 (*)	complete	chromosome	CP015350.1	Cry13	nematodes
Bt MYBT18246_2 (*)	complete	chromosome	CP015350.1	Cry13	nematodes
Bt MYBT18246_3 (*)	complete	chromosome	CP015350.1	Cry13	nematodes
Bt Sbt003_1 (**)	draft	contig 2	AMYJ01000002.1	Cry65	not yet known
Bt Sbt003_2 (**)	draft	contig 14	AMYJ01000014.1	Cry5	nematodes
Bt Sbt003_3 (**)	draft	contig 103	AMYJ01000103.1	Cry5	nematodes
Bt YBT-1518 (***)	complete	chromosome	CP005935.1	Cry5	nematodes

(*) [13], (**) [17], (***) [12], (+) Direct submission

A closer evaluation of the complete prophage-associated *cry* cassettes from either complete or draft genomes is depicted in **Figure 2**. The comparison revealed that most of the cassettes identified within draft genomes are located near the contig's end. As a result, the downstream prophage region is incomplete in most of these cases and is even missing in Bt AFS063736. Yet, further characterization of the identified prophage-associated *cry* cassettes exposed an interesting feature. In five of the eight draft genome cases, *cry5* genes were detected, these cassettes correspond with that of Bt YBT-1518. Surprisingly, *cry65* genes were identified in the prophage-associated *cry* cassettes of the three remaining draft genomes. Similar to the three instances in Bt MYBT18246, the *cry65* genes were upstream to a coding sequence of a pesticidal C-terminal region as part of the split toxin variant. Overall, twelve complete prophage-associated *cry* cassettes that include either the gene *cry13*, *cry5*, or *cry65* were identified. It is of notice that Bt Sbt003 and Bc VD154, both present multiple cassettes variants with either a *cry5* or a *cry65* gene. No other insertion element was detected in the vicinity of the prophage-associated cassette, except for one case in Bt MYBT18246, which presented flanking insertion sequences. Lastly, to predict the genomic localization of the eight draft genome cassettes, a contig rearrangement analysis using a complete reference genome and the Mauve application [18] was conducted (**Figure S1**). Despite the limited reliability when rearranging data sets of draft genomes, our results support a putative chromosomal localization for all identified draft genome cases.

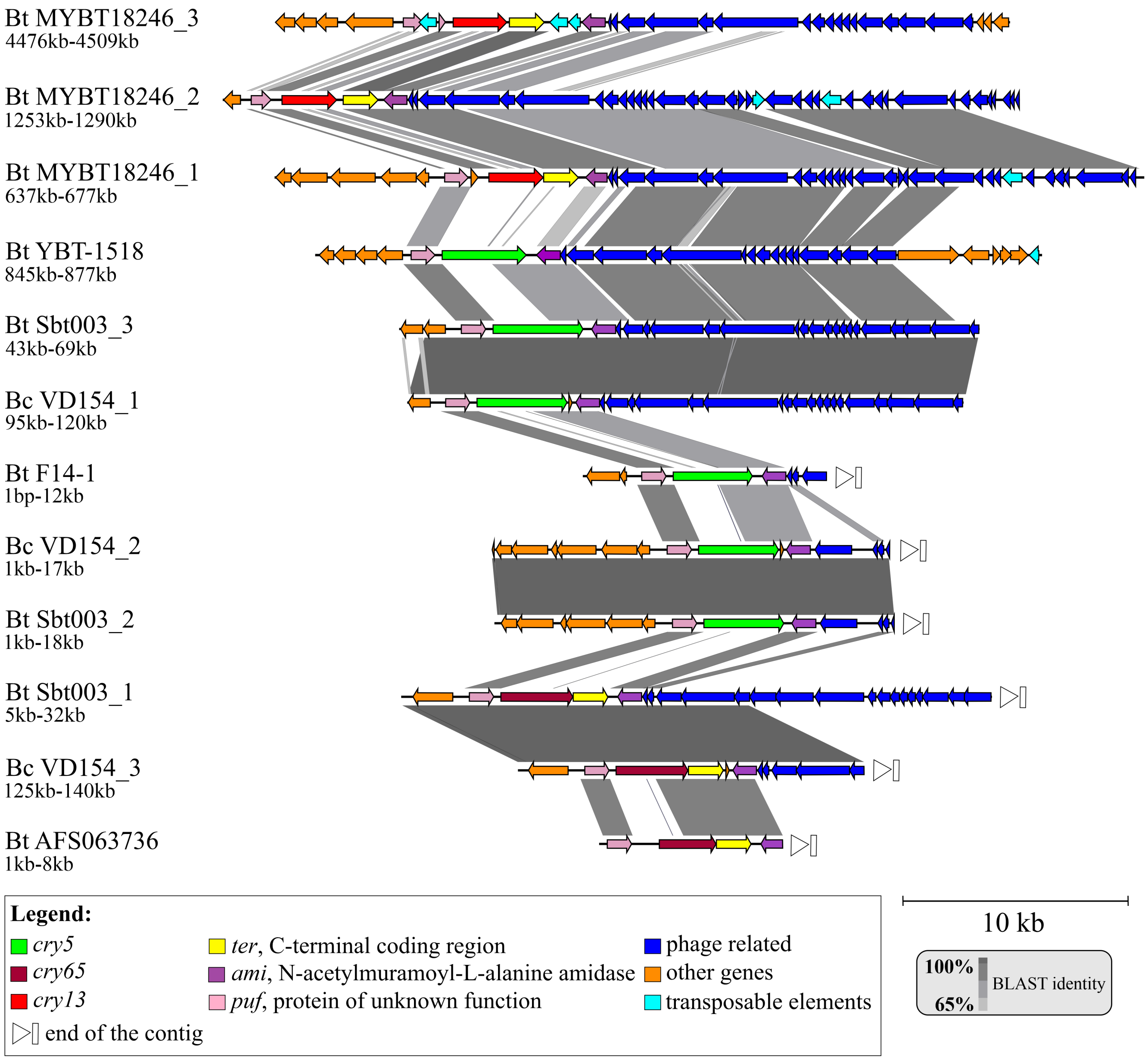


Figure 2. Genomic loci comparison of all identified prophage-associated *cry* cassettes. Cassettes found within the complete genomes of *B. thuringiensis* MYBT18246 and *B. thuringiensis* YBT-1518 are aligned using Easyfig [19] with the additional draft genome cassettes retrieved by BLASTn [16]. Strain names and chromosomal coordinates appear next to each row. Partial prophage regions caused by the location of the cassette near the end of the contig are directly indicated.

2.2. *Characterization of the novel prophage-associated pesticidal cassette*

A protein-based phylogenetic analysis of each cassette component resulted in the maximum-likelihood trees shown in **Figure 3**. The proteins of unknown function and N-acetylmuramoyl-L-alanine amidases encoded by the *puf* and *ami* genes of the prophage-associated *cry* cassette are highly conserved. **Figure 3A** and **Figure 3B** depicts maximum-likelihood trees of these proteins. A further sequence evaluation showed that in both cases the protein sequences were conserved, with a corresponding average of 91% and 87% sequence similarity between the proteins of unknown function and the amidases, respectively. **Figure 3C** and **Figure 3D** depicts the maximum-likelihood trees of the encoded pesticidal toxins and their either integrated or separated C-terminal regions, respectively. The protein sequences are distributed into three distinct groups according to their three identified primary toxin ranks. Additional sequence evaluation showed that as opposed to the two previous trees, here, the sequences are considerably more diverse. For example, the average sequence similarity within each of the pesticidal groups, Cry5, Cry13, and Cry65, is 65%, 80%, and 76%, respectively. When calculating the sequence similarity between the three pesticidal groups, the average percentage was even lower, with a value of 31%.

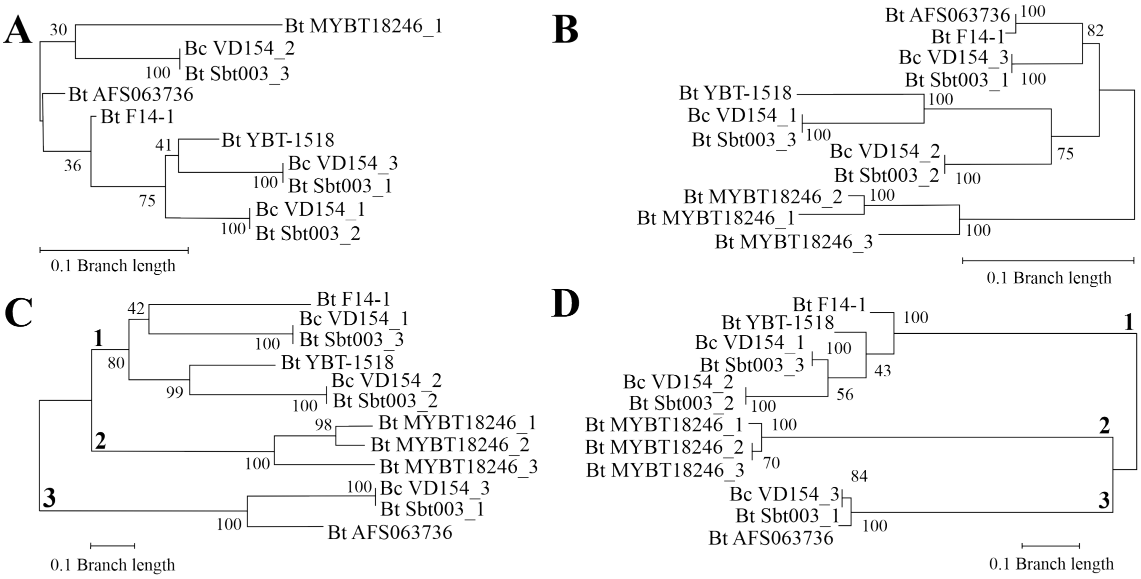


Figure 3. Protein-based maximum-likelihood trees of each component within the identified prophage-associated *cry* cassettes. **(A)** Proteins of unknown function. **(B)** N-acetylmuramoyl-L-alanine amidases. **(C)** Cry toxins. Branches 1, 2, and 3 create a division between the sequences of Cry5, Cry13, and Cry65, respectively. **(D)** Pesticidal C-terminal regions. Branch 1 groups all of the Cry5-integrated C-terminal regions. Branches 2 and 3 refer to the independent C-terminal coding sequences that appear downstream to the *cry13* and *cry65* pesticidal genes. For this figure, all of the corresponding accession numbers can be found in **Table S3**.

Domain structure analysis of the Cry5, Cry65, and Cry13 toxins encoded within the identified prophage-associated *cry* cassettes was performed using CDvist [20] to further understand their sequence variability (**Figure 4**). Here, a positive control in the form of a recognized plasmid-encoded Cry5 sequence (Cry5Aa1 from Bt serovar *darmstadiensis*; accession AAA67694.1) was used to emphasize the domain order within three-domain Cry proteins. In this analysis, the varying sequence lengths and domain configurations highlight the diversity among the Cry toxins compared to other proteins that belong to the same cassette. Here, Cry65 and Cry13 consist of a core region that includes the N-terminal domain (InterPro entry IPR005639), the central domain (InterPro entry IPR001178), and the C-terminal domain (InterPro entry IPR005638) of a typical three-domain Cry toxin. The two sequence types represent the short toxin variant that is paired with the independent C-terminal region, which contains the characterizing domain V (InterPro entry IPR041587). Cry5 represents the long toxin variant that consists of both the core region and the C-terminal region.

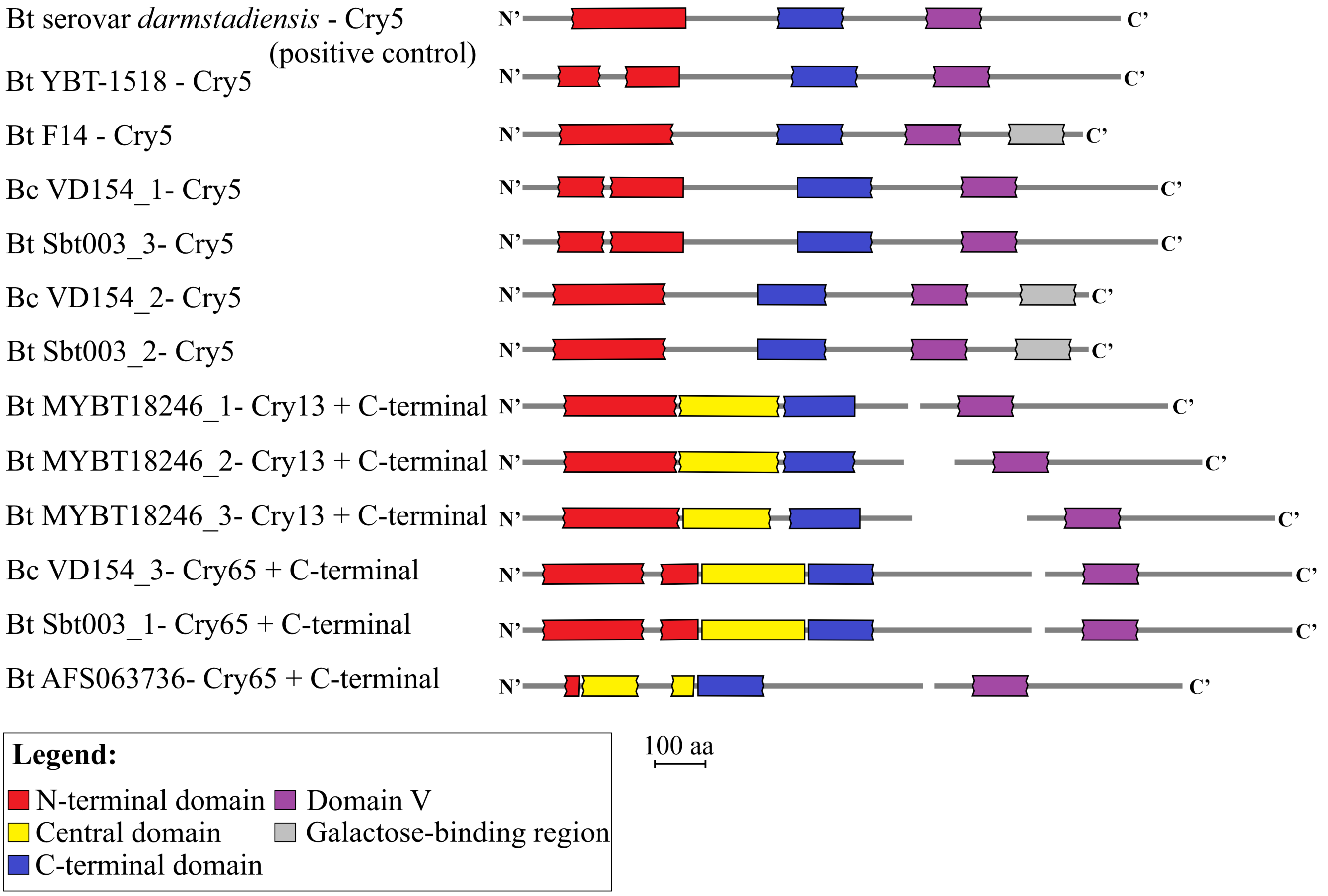

Bt serovar *darmstadiensis* - Cry5
(positive control)
Bt YBT-1518 - Cry5
Bt F14 - Cry5
Bc VD154_1- Cry5
Bt Sbt003_3- Cry5
Bc VD154_2- Cry5
Bt Sbt003_2- Cry5
Bt MYBT18246_1- Cry13 + C-terminal
Bt MYBT18246_2- Cry13 + C-terminal
Bt MYBT18246_3- Cry13 + C-terminal
Bc VD154_3- Cry65 + C-terminal
Bt Sbt003_1- Cry65 + C-terminal
Bt AFS063736- Cry65 + C-terminal
N'
C'
100 aa
Legend:
🟥 N-terminal domain
🟨 Central domain
🟦 C-terminal domain
🟪 Domain V
⬜ Galactose-binding region

Figure 4. Domain overview of aligned Cry toxins encoded within the identified prophage-associated *cry* cassettes: All identified Cry proteins correspond to three-domain toxins. Cry65 and Cry13 sequences are encoded by short toxin gene variants followed by an independent pesticidal C-terminal gene with the typical domain V. Cry5 includes both the core and the C-termianl regions in one protein. Note that the distances between proteins correspond to the gap length between the coding sequences on the genomic level. Plasmid-encoded Cry5 from Bt serovar *darmstadiensis* was added for comparison. Partial or interrupted domain matches are indicated by borders with a wavy appearance. More details and sequence accessions can be found in **Table S3**.

2.3. Evaluation of prophage regions

All of the identified cassette-associated prophage regions were evaluated using Prophage Hunter [21] and PHASTER [22]. Both assigned the prophages to the family *Siphoviridae*. Additionally, the evaluations included the *ami* gene as part of predicted prophage regions. Further analysis of the two complete chromosomes belonging to Bt MYBT18246 and Bt YBT-1518 determined that even though several complete prophage regions exist in both chromosomes, all of the cassette-associated prophage regions are either inconclusively complete or incomplete prophages (**Figure S2** and **Table S4**). **Figure 5** compares the prophage regions of *B. thuringiensis* MYBT18246 and *B. thuringiensis* YBT-1518 with three *Siphoviridae* genomes that were suggested as closely-related phage genomes. The comparison shows that each of the presented *Siphoviridae* phages, vB_BtS_BMBBtp1 (accession KT852578.1), vB_BtS_BMBBtp15 (accession KX190835.1), and vB_BtS_BMBBtp16 (accession KT372714.1), encodes an amidase gene. Further sequence evaluation revealed that these share an average of 72% sequence similarity with the amidases of the prophage-associated *cry* cassettes.

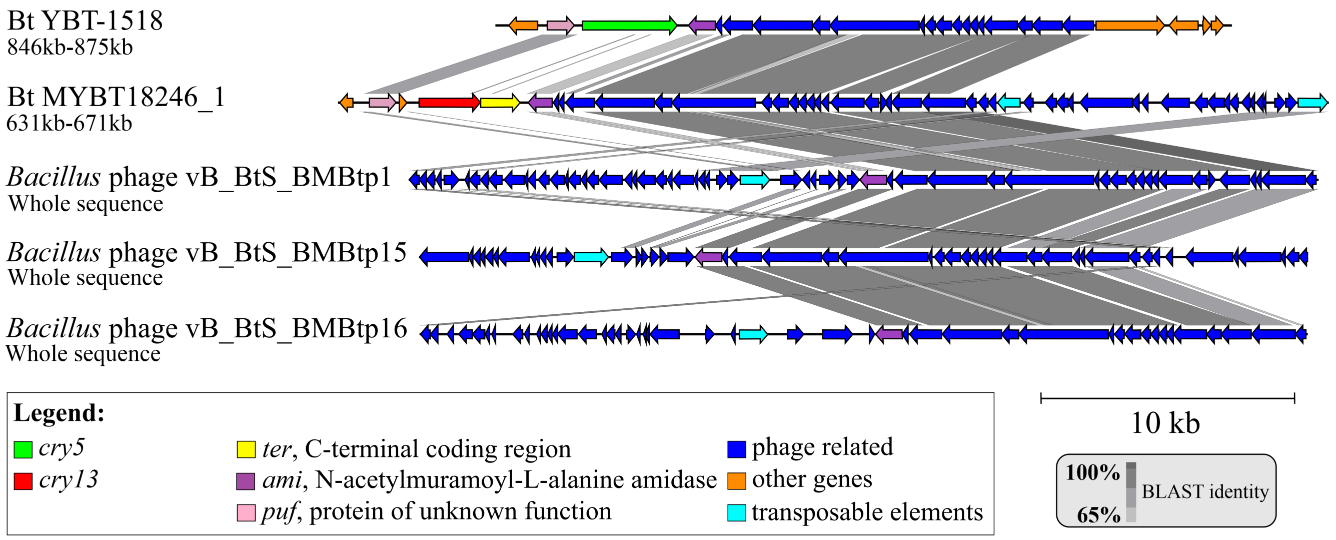

Bt YBT-1518
846kb-875kb
Bt MYBT18246_1
631kb-671kb
Bacillus phage vB_BtS_BMBtp1
Whole sequence
Bacillus phage vB_BtS_BMBtp15
Whole sequence
Bacillus phage vB_BtS_BMBtp16
Whole sequence
Legend:
🟩 *cry5*
🟥 *cry13*
🟨 *ter*, C-terminal coding region
🟪 *ami*, N-acetylmuramoyl-L-alanine amidase
🟪 *puf*, protein of unknown function
🟦 phage related
🟧 other genes
🟦 transposable elements
10 kb
100%
65%
BLAST identity

Figure 5. Comparison of the identified prophage regions with closest known phages. Two recognized cassette-associated prophage sequences from Bt YBT-1518 and Bt MYBT18246 are compared using Easyfig [19] with three *Siphoviridae* genomes suggested by the utilized prediction tools as closely-related phage genomes.

3. Discussion

While investigating the genetic environment of chromosomally-encoded *cry* genes in complete and draft *B. thuringiensis* genomes, twelve instances of a conserved three-component genetic pattern, which was defined as the prophage-associated *cry* cassette, were identified. While chromosomal *cry* genes were previously reported in a handful of instances [12,13], they were, to our knowledge, never aligned and compared. Among those, only [13] mentioned the proximity of the *cry* genes to prophage regions. In this study, a comparative genomic analysis of chromosomally-encoded *cry* toxins resulted in the identification of a novel cassette and confirmed that an association with a prophage has evolved for three different toxin types.

The first conserved element within the prophage-associated *cry* cassette is a *puf* gene. The original annotations of the *puf* genes, which were extracted from GenBank, varied from 'replication protein' to 'cytosolic protein' and 'hypothetical protein'. On the amino acid level, a probable helix-turn-helix motif was detected when conducting secondary structure predictions, which suggest a DNA binding function and a possible role in gene expression. Additionally, the high similarity among the *puf* sequences may indicate the importance of this encoded protein. Nevertheless, to our knowledge, no functional characterization is available for the encoded protein and its exact role remains unknown. The second conserved element is the *ami* gene, which encodes an N-acetylmuramoyl-L-alanine amidase. *B. thuringiensis* amidases are expressed during the late sporulation phase and are involved in lysis of the mother cell, which in turn results in spore and crystallized toxin release [23,24]. Interestingly, in addition to their existence in various bacterial species, amidase genes were observed in bacteriophage genomes [25,26]. Moreover, [27] identified a plasmid-encoded *cry1* cassette in *B. thuringiensis*. Their identified cassette contains an *ami* sequence that was assumed to have a phage origin. Nevertheless, according to our research, the amino acid sequence of the amidases from the prophage-associated cassette shares an average sequence similarity of 36% with the amidases that were recognized in the aforementioned study. The relatively low similarity between these amidases suggests two distinct groups of sequences that derive from different phage origins.

The third and most important component of the prophage-associated *cry* cassette is the pesticidal gene that is located between the *puf* and *ami* genes. Among the identified cassettes, three different types of *cry* genes (*cry5*, *cry13*, and *cry65*), which belong to the three-domain toxin family, were found. Cry65 has been described as being active against cancer cells but currently has an unknown natural host [28]. In contrast, Cry5 and Cry13 both target nematodes [29]. This case of sequence diversity and host specificity is of special interest when considering the conserved surroundings and the localization of the identified toxin genes. In addition to the toxin-encoding gene, a separate pesticidal C-terminal coding sequence (*ter*), which was found to be necessary for crystallization and stabilization of the toxin [28,30], is located directly downstream to the short chromosomally-encoded gene variants, *cry65* and *cry13*. The gene *ter*, also commonly described as *orf2*, shows high homology with the C-terminal coding region of the longer *cry* gene variants, among which is the chromosomally-encoded *cry5* gene. Without further experiments, the question regards the reason for maintaining two variants and which toxin variant is preferable remains open. Interestingly, other examples of such split variants have been previously recognized and discussed in terms of evolutionary adaptation [31,32]. Nevertheless, whether the short variant resulted from a split in the toxin coding sequence or the long variant was derived from a gene fusion event of independently acquired *cry* and *orf2* genes remains an open question. Partial genes have been proposed as an evolutionary strategy of *B. thuringiensis* to diversify its toxin armory and adapt to different hosts [28,33]. Furthermore, [34] showed that an artificial recombination of three-domain *cry* genes resulted in a chimeric Cry4Ba/Cry1Ac toxin that exhibites an 238-fold increased toxicity against *Culex pipiens* [34]. This evidence indicates that the occurrence of recombination events, like those which led to the *cry* gene diversification within the prophage-associated *cry* cassette, hold the potential to create stronger and thus better toxins.

It has been shown that major clades of *B. thuringiensis* accumulate virulence factors that are active against certain targets and specialize in creating a specific host range [35]. Such repertoire can have synergistic effects, which increases the toxicity and assists in counteracting host resistance mechanisms [36]. Two examples are Bt YBT-1518 that carries the plasmids-encoded *cry55Aa1*, *cry6Aa2*, and the chromosome-encoded *cry5Ba2* gene, all of which have reported nematocidal activity, and Bt serovar *israelensis*, which produces a crystal composed of six different toxins that have synergistic effects towards dipteran pests [37,38]. Alternatively, some Bt strains produce pesticidal proteins that are active against invertebrates of different orders. Such diversity within one strain could be the outcome of co-evolutionary host alternation mechanisms [35]. Many environments in which Bt thrives consist of many potential hosts, which constantly develop resistance strategies against its virulence factors. *B. thuringiensis*, in turn, might embrace a host switching strategy to take advantage of the least protected pest [35]. An interesting example from this study is found in Bt Sbt003 and *B. cereus* VD154, which harbor two variants of nematocidal *cry5* genes within their prophage-associated cassette. These encoded-toxins might synergize to avoid host resistance [39]. Furthermore, both of these Bt strains carry a *cry65* gene toxic against an unknown natural host. This case reveals that the same prophage-associated cassette harbors different toxin genes that may target diverse hosts. Nematodes might represent a point of interaction in which nematocidal strains can encounter dipteran strains, and thus perform the horizontal gene exchange required for host alternation strategies, either by transconjugative plasmids [38] or perhaps by phages.

Re-occurring proximity to a prophage region is another exceptional feature of the identified cassettes. The comparison of the cassette-associated prophage regions with known Bt phages **Figure 5** suggests a relation. Nevertheless, phage prediction presents a challenge for bioinformatics since the assignment of phage families strongly rely on the narrow subset of known information within the extremely diverse world of *Bacillus* phages [40]. The phage prediction tool PHASTER [22] described the cassette-adjacent prophages as incomplete or inconclusively complete *Siphoviridae*-like prophages. The prophages may have become incomplete after their chromosomal integration as a result of the bacterial antiphage defense mechanism [41]. Furthermore, incompleteness does not mean that the prophages cannot still become active under appropriate conditions and selective pressure [42]. Studies have shown that in *Enterococcus faecalis* and *Staphylococcus aureus* prophages that appear impaired can still replicate by using a minimal amount of essential genes or by taking advantage of other complete prophages, which are found within the genome [43]. Consequently, experimental data is required to assess and confirm the predicted status of the found prophages.

Our results suggest that the components of the prophage-associated *cry* cassette have evolved from a common ancestor. Notably, the high conservation in the *puf* and *ami* genes stands in contrast to the diversity of the pesticidal genes that they surround. [30] indicated plasmids, transposition, and recombination as being the three major factors behind the evolution of Bt toxins. The first factor does not apply in the case of the prophage-associated *cry* cassette since the cassette has chromosomal localization. The second factor does not apply since, except for one case in Bt MYBT18246, the only annotated mobile element is the prophage itself. Thus, recombination remains a possible mechanism to generate genetic variability within a conserved genomic arrangement. Three-domain Cry toxins contain blocks of shared homology [3,44], which may serve as potential points for homologous recombination. Yet, open questions remain regarding a possible transfer mechanism between different strains within the species. The idea of activated phages that serve as vectors for *cry* genes is very appealing and is found in agreement with the phenomenon of lysogenic conversion, which is often observed in *Vibrio cholera*, *E. faecalis*, and *S. aureus*[15,43]. To determine whether the cassette-associate prophages play a role in mobilization of *cry* genes between Bt strains, experimental evidence of phage particles carrying the identified *cry* cassette is crucial.

To conclude, a conserved arrangement of a prophage-associated *cry* cassette consisting of three components was discovered in *B. thuringiensis* genomes. The chromosomally-encoded cassette consists of highly conserved non-pesticidal components and a variable *cry* toxins. Comparative genomics indicates a contribution of gene exchange by recombination to the evolution of Cry toxins. The discovery of a chromosomal prophage-associated *cry* cassette opens the door to further research on non-plasmid vectors for the horizontal acquisition of *cry* genes by *B. thuringiensis*. Our discoveries push towards further understanding of host driven adaptation mechanisms that might be applied for new biopesticide development approaches.

4. Materials and Methods

4.1. Establishing a genome collection and comparison of the cry-encoding loci

For the comparison of *cry*-encoding regions, all complete *B. thuringiensis* genomes (61 in total) were downloaded as GenBank files from the NCBI collection (www.ncbi.nlm.nih.gov/genome; accessed last on Dec. 2020). To generate a uniform and comparable dataset, all genomes were re-annotated by the NCBI Prokaryotic Genome Annotation Pipeline (PGAP) [45], according to the recommended parameters. To further identify and compare *cry* genes within the re-annotated genomes, an in-house developed software, IDOPS (Díaz-Valerio, unpublished), was used. Briefly, IDOPS (Identification of Pesticidal Sequences) detects pesticidal sequences according to the BPPRC classification, extracts the genetic environment of candidate genes, and compares them by generating a visualization of the common loci using EasyFig v.2.2.2 [19]. IDOPS detection is based on a curated collection of high-quality profile hidden Markov models (model database is provided in the supplementary). Finally, a default BLASTp [16] search was done to verify the rank of each identified toxin (see **Table S1** for rank assignments).

4.2. Extended search for pesticidal cassettes associated with prophages

Comparative analysis of complete genomes identified *cry* coding sequences surrounded by a repeating set of genes and located upstream to a prophage region. This genetic arrangement was defined as a prophage-associated *cry* cassette. To investigate whether this cassette also represents a re-occurring pattern within draft genomes of Bt, a

BLASTn [16] search was conducted. The search was done against the RefSeq database and was restricted to the *B. cereus* group (tax-id: 86661). As query, the genes surrounding the chromosomal *cry* gene were used, while the *cry* gene itself was masked with 'N' characters. The search was optimized for more dissimilar sequences by using a discontinuous megablast. This was done to expand the search beyond the already identified *cry* genes within the Bt YBT-1518 and Bt MYBT18246 cassettes. The search returned variants of the identified cassette within draft genomes of Bt. The GenBank files of the matching genomic regions were evaluated by a default IDOPS search (Díaz-Valerio, unpublished) to assess the presence of pesticidal sequences. Finally, to determine the putative genomic localization (i.e. chromosomal or plasmid-encoded) of each matching sequence, a rearrangement of each cassette-encoding draft genome and a comparison with the reference Bt MYBT18246 genome were performed using MAUVE v.2.3.1 [18].

4.3. Characterization of the novel prophage-associated cry cassette

Protein translations of each gene within the identified pesticidal cassettes were obtained directly from the GenBank files. The protein sequences were aligned by ClustalW2 v.2.1 [46] using the default settings and represented in maximum-likelihood trees with a 1000-replicate bootstrap by using MEGA v.10.0.5 [47]. Analysis of sequence similarities was carried out using the program Ident and Sim [48] from the sequence manipulation suite at the bioinformatics.org website (bioinformatics.org/sms2/ident_sim.html; accessed on Dec. 2020). Protein domain analysis of the cassete-encoded pesticidal sequences was done using CDvist [20] with the parameter 'HHsearch probability' set to be 60% or more (cdvist.zhulinlab.org; accessed last on Jan. 2021). Additional secondary structure predictions for the protein of unknown function were performed using the program Quick2D [49] from the bioinformatics toolkit of the Max Planck Institute, Germany (toolkit.tuebingen.mpg.de/tools/quick2d; accessed last on Jan. 2021). Finally, Prophage Hunter [21] and PHASTER [22] were used to ascertain the presence of prophage regions in proximity to the identified pesticidal cassettes (pro-hunter.genomics.cn, phaster.ca; both accessed last on Dec. 2020), and IS-finder [50] was used to detect insertion sequences in the 5000bp region upstream and downstream to each identified cassette (isfinder.biotoul.fr; accessed last on Jan. 2021).

Supplementary Materials: The following are available online at https://www.mdpi.com/article/10.3390/toxins1010000/s1, Figure S1: title, Table S1: title, Video S1: title.

Author Contributions: For research articles with several authors, a short paragraph specifying their individual contributions must be provided. The following statements should be used "Conceptualization, X.X. and Y.Y.; methodology, X.X.; software, X.X.; validation, X.X., Y.Y. and Z.Z.; formal analysis, X.X.; investigation, X.X.; resources, X.X.; data curation, X.X.; writing—original draft preparation, X.X.; writing—review and editing, X.X.; visualization, X.X.; supervision, X.X.; project administration, X.X.; funding acquisition, Y.Y. All authors have read and agreed to the published version of the manuscript.", please turn to the CRediT taxonomy for the term explanation. Authorship must be limited to those who have contributed substantially to the work reported.

ALH performed research, generated alignments, analyzed the data, wrote the manuscript, SDV performed research, analyzed the data, wrote the manuscript, JH discovered the initial phage-associated toxins, wrote the manuscript, RSS processed data, generated alignments, wrote the manuscript, HL designed the study, evaluated the data, wrote the manuscript

Funding: We acknowledge the support by the Open Access Publication Funds of the Göttingen University, Germany. The funders had no role in the design of the study; in collection, analyses or interpretation of the data; in the writing of the manuscript or in the decision to publish the results.

Acknowledgments: We thank Tarek Morsi for the expert IT support with the G2L servers. We thank as well Prof. Rolf Daniel for hosting and supporting us, and giving us unlimited access to the G2L-IT infrastructure.

343 **Conflicts of Interest:** The authors declare no conflict of interest.

References

1. Ehling-Schulz, M.; Lereclus, D.; Koehler, T.M. The *Bacillus cereus* Group: *Bacillus* Species with Pathogenic Potential. *Microbiology Spectrum* **2019**, *7*. doi:10.1128/microbiolspec.gpp3-0032-2018.
2. Mendoza-Almanza, G.; Esparza-Ibarra, E.L.; Ayala-Luján, J.L.; Mercado-Reyes, M.; Godina-González, S.; Hernández-Barrales, M.; Olmos-Soto, J. The cytocidal spectrum of *Bacillus thuringiensis* toxins: From insects to human cancer cells. *Toxins* **2020**, *12*, 301. doi:10.3390/toxins12050301.
3. Palma, L.; Muñoz, D.; Berry, C.; Murillo, J.; Caballero, P.; Caballero, P. *Bacillus thuringiensis* toxins: An overview of their biocidal activity. *Toxins* **2014**, *6*, 3296–3325. doi:10.3390/toxins6123296.
4. Crickmore, N.; Berry, C.; Panneerselvam, S.; Mishra, R.; Connor, T.R.; Bonning, B.C. A structure-based nomenclature for *Bacillus thuringiensis* and other bacteria-derived pesticidal proteins. *Journal of Invertebrate Pathology* **2020**, p. 107438. doi:10.1016/j.jip.2020.107438.
5. Bravo, A.; Likitvivatanavong, S.; Gill, S.S.; Soberón, M. *Bacillus thuringiensis*: A story of a successful bioinsecticide. *Insect Biochemistry and Molecular Biology* **2011**, *41*, 423–431. doi:10.1016/j.ibmb.2011.02.006.
6. Azizoglu, U.; Jouzani, G.S.; Yilmaz, N.; Baz, E.; Ozkok, D. Genetically modified entomopathogenic bacteria, recent developments, benefits and impacts: A review. *Science of the Total Environment* **2020**, *734*, 139169. doi:10.1016/j.scitotenv.2020.139169.
7. Huang, Y.J.S.; Higgs, S.; Vanlandingham, D.L. Biological control strategies for mosquito vectors of arboviruses. *Insects* **2017**, *8*, 21. doi:10.3390/insects8010021.
8. George, Z.; Crickmore, N. *Bacillus thuringiensis* applications in agriculture. In *Bacillus thuringiensis biotechnology*; Springer, 2012; pp. 19–39.
9. Patiño-Navarrete, R.; Sanchis, V. Evolutionary processes and environmental factors underlying the genetic diversity and lifestyles of *Bacillus cereus* group bacteria. *Research in microbiology* **2017**, *168*, 309–318.
10. Lechuga, A.; Lood, C.; Salas, M.; van Noort, V.; Lavigne, R.; Redrejo-Rodríguez, M. Completed genomic sequence of *Bacillus thuringiensis* HER1410 reveals a cry-containing chromosome, two megaplasmids, and an integrative plasmidial prophage. *G3: Genes, Genomes, Genetics* **2020**, *10*, 2927–2939. doi:10.1534/g3.120.401361.
11. Makart, L.; Commans, F.; Gillis, A.; Mahillon, J. Horizontal transfer of chromosomal markers mediated by the large conjugative plasmid pXO16 from *Bacillus thuringiensis* serovar *israelensis*. *Plasmid* **2017**, *91*, 76–81. doi:10.1016/j.plasmid.2017.04.001.
12. Wang, P.; Zhang, C.; Guo, M.; Guo, S.; Zhu, Y.; Zheng, J.; Zhu, L.; Ruan, L.; Peng, D.; Sun, M. Complete genome sequence of *Bacillus thuringiensis* YBT-1518, a typical strain with high toxicity to nematodes. *Journal of Biotechnology* **2014**, *171*, 1–2. doi:10.1016/j.jbiotec.2013.11.023.
13. Hollensteiner, J.; Poehlein, A.; Spröer, C.; Bunk, B.; Sheppard, A.E.; Rosentstiel, P.; Schulenburg, H.; Liesegang, H. Complete Genome sequence of the nematicidal *Bacillus thuringiensis* MYBT18246. *Standards in Genomic Sciences* **2017**, *12*, 48. doi:10.1186/s40793-017-0259-x.
14. Gillis, A.; Mahillon, J. Phages Preying on *Bacillus anthracis*, *Bacillus cereus*, and *Bacillus thuringiensis*: Past, Present and Future. *Viruses* **2014**, *6*, 2623–2672. doi:10.3390/v6072623.
15. Penadés, J.R.; Christie, G.E. The Phage-Inducible Chromosomal Islands: A Family of Highly Evolved Molecular Parasites. *Annual Review of Virology* **2015**, *2*, 181–201. PMID: 26958912, doi:10.1146/annurev-virology-031413-085446.
16. Camacho, C.; Coulouris, G.; Avagyan, V.; Ma, N.; Papadopoulos, J.; Bealer, K.; Madden, T.L. BLAST+: Architecture and applications. *BMC Bioinformatics* **2009**, *10*, 421. doi:10.1186/1471-2105-10-421.
17. Liu, Y.; Ye, W.; Zheng, J.; Fang, L.; Peng, D.; Ruan, L.; Sun, M. High-quality draft genome sequence of nematocidal *Bacillus thuringiensis* Sbt003. *Standards in Genomic Sciences* **2015**, *9*, 624–631. doi:10.4056/sigs.4738557.
18. Darling, A.C.; Mau, B.; Blattner, F.R.; Perna, N.T. Mauve: Multiple alignment of conserved genomic sequence with rearrangements. *Genome Research* **2004**, *14*, 1394–1403. doi:10.1101/gr.2289704.
19. Sullivan, M.J.; Petty, N.K.; Beatson, S.A. Easyfig: A genome comparison visualizer. *Bioinformatics* **2011**, *27*, 1009–1010. doi:10.1093/bioinformatics/btr039.
20. Adebali, O.; Ortega, D.R.; Zhulin, I.B. CDvist: A webserver for identification and visualization of conserved domains in protein sequences. *Bioinformatics* **2015**, *31*, 1475–1477. doi:10.1093/bioinformatics/btu836.
21. Song, W.; Sun, H.X.; Zhang, C.; Cheng, L.; Peng, Y.; Deng, Z.; Wang, D.; Wang, Y.; Hu, M.; Liu, W.; Yang, H.; Shen, Y.; Li, J.; You, L.; Xiao, M. Prophage Hunter: an integrative hunting tool for active prophages. *Nucleic Acids Research* **2019**, *47*, W74–W80. doi:10.1093/nar/gkz380.
22. Arndt, D.; Grant, J.R.; Marcu, A.; Sajed, T.; Pon, A.; Liang, Y.; Wishart, D.S. PHASTER: a better, faster version of the PHAST phage search tool. *Nucleic acids research* **2016**, *44*, W16–W21. doi:10.1093/nar/gkw387.
23. Yang, J.; Peng, Q.; Chen, Z.; Deng, C.; Shu, C.; Zhang, J.; Huang, D.; Song, F. Transcriptional regulation and characteristics of a novel N-acetylmuramoyl-L-alanine amidase gene involved in *Bacillus thuringiensis* mother cell lysis. *Journal of bacteriology* **2013**, *195*, 2887–2897.
24. Chen, X.; Gao, T.; Peng, Q.; Zhang, J.; Chai, Y.; Song, F. Novel cell wall hydrolase CwlC from *Bacillus thuringiensis* is essential for mother cell lysis. *Applied and environmental microbiology* **2018**, *84*.

25. Sass, P.; Bierbaum, G. Lytic Activity of Recombinant Bacteriophage ϕ11 and ϕ12 Endolysins on Whole Cells and Biofilms of *Staphylococcus aureus*. *Applied and Environmental Microbiology* **2006**, *73*, 347–352. doi:10.1128/aem.01616-06.
26. Son, B.; Kong, M.; Ryu, S. The Auxiliary Role of the Amidase Domain in Cell Wall Binding and Exolytic Activity of Staphylococcal Phage Endolysins. *Viruses* **2018**, *10*, 284. doi:10.3390/v10060284.
27. Fiedoruk, K.; Daniluk, T.; Mahillon, J.; Leszczynska, K.; Swiecicka, I. Genetic environment of *cry1* genes indicates their common origin. *Genome Biology and Evolution* **2017**, *9*, 2265–2275. doi:10.1093/gbe/evx165.
28. Peng, D.H.; Pang, C.Y.; Wu, H.; Huang, Q.; Zheng, J.S.; Sun, M. The expression and crystallization of Cry65Aa require two C-termini, revealing a novel evolutionary strategy of *Bacillus thuringiensis* Cry proteins. *Scientific Reports* **2015**, *5*. doi:10.1038/srep08291.
29. Jouzani, G.S.; Valijanian, E.; Sharafi, R. *Bacillus thuringiensis*: a successful insecticide with new environmental features and tidings. *Applied microbiology and biotechnology* **2017**, *101*, 2691–2711.
30. de Maagd, R.A.; Bravo, A.; Berry, C.; Crickmore, N.; Schnepf, H.E. Structure, diversity, and evolution of protein toxins from spore-forming entomopathogenic bacteria. *Annual review of genetics* **2003**, *37*, 409–433.
31. Strittmatter, A.W.; Liesegang, H.; Rabus, R.; Decker, I.; Amann, J.; Andres, S.; Henne, A.; Fricke, W.F.; Martinez-Arias, R.; Bartels, D.; Goesmann, A.; Krause, L.; Pühler, A.; Klenk, H.P.; Richter, M.; Schüler, M.; Glöckner, F.O.; Meyerdierks, A.; Gottschalk, G.; Amann, R. Genome sequence of *Desulfobacterium autotrophicum* HRM2, a marine sulfate reducer oxidizing organic carbon completely to carbon dioxide. *Environmental Microbiology* **2009**, *11*, 1038–1055. doi:10.1111/j.1462-2920.2008.01825.x.
32. Berger, M.; Brock, N.L.; Liesegang, H.; Dogs, M.; Preuth, I.; Simon, M.; Dickschat, J.S.; Brinkhoff, T. Genetic Analysis of the Upper Phenylacetate Catabolic Pathway in the Production of Tropodithietic Acid by *Phaeobacter gallaeciensis*. *Applied and Environmental Microbiology* **2012**, *78*, 3539–3551. doi:10.1128/aem.07657-11.
33. Sajid, M.; Geng, C.; Li, M.; Wang, Y.; Liu, H.; Zheng, J.; Peng, D.; Sun, M. Whole-genome analysis of *Bacillus thuringiensis* revealing partial genes as a source of novel Cry toxins. *Applied and environmental microbiology* **2018**, *84*.
34. Zghal, R.Z.; Elleuch, J.; Ali, M.B.; Darriet, F.; Rebaï, A.; Chandre, F.; Jaoua, S.; Tounsi, S. Towards novel Cry toxins with enhanced toxicity/broader: a new chimeric Cry4Ba / Cry1Ac toxin. *Applied Microbiology and Biotechnology* **2016**, *101*, 113–122. doi:10.1007/s00253-016-7766-3.
35. Zheng, J.; Gao, Q.; Liu, L.; Liu, H.; Wang, Y.; Peng, D.; Ruan, L.; Raymond, B.; Sun, M. Comparative genomics of *Bacillus thuringiensis* reveals a path to specialized exploitation of multiple invertebrate hosts. *MBio* **2017**, *8*.
36. Fayad, N.; Kambris, Z.; El Chamy, L.; Mahillon, J.; Kallassy Awad, M. A novel anti-dipteran *Bacillus thuringiensis* strain: Unusual Cry toxin genes in a highly dynamic plasmid environment. *Applied and Environmental Microbiology* **2020**. doi:10.1128/aem.02294-20.
37. Yu, Z.; Luo, H.; Xiong, J.; Zhou, Q.; Xia, L.; Sun, M.; Li, L.; Yu, Z. *Bacillus thuringiensis* Cry6A exhibits nematicidal activity to *Caenorhabditis elegans bre* mutants and synergistic activity with Cry5B to *C. elegans*. *Letters in applied microbiology* **2014**, *58*, 511–519.
38. Ruan, L.; Crickmore, N.; Peng, D.; Sun, M. Are nematodes a missing link in the confounded ecology of the entomopathogen *Bacillus thuringiensis*? *Trends in microbiology* **2015**, *23*, 341–346.
39. Geng, C.; Liu, Y.; Li, M.; Tang, Z.; Muhammad, S.; Zheng, J.; Wan, D.; Peng, D.; Ruan, L.; Sun, M. Dissimilar crystal proteins Cry5Ca1 and Cry5Da1 synergistically act against *Meloidogyne incognita* and delay Cry5Ba-based nematode resistance. *Applied and environmental microbiology* **2017**, *83*.
40. Grose, J.H.; Jensen, G.L.; Burnett, S.H.; Breakwell, D.P. Genomic comparison of 93 *Bacillus* phages reveals 12 clusters, 14 singletons and remarkable diversity. *BMC Genomics* **2014**, *15*, 855. doi:10.1186/1471-2164-15-855.
41. Doron, S.; Melamed, S.; Ofir, G.; Leavitt, A.; Lopatina, A.; Keren, M.; Amitai, G.; Sorek, R. Systematic discovery of antiphage defense systems in the microbial pangenome. *Science* **2018**, *359*, eaar4120. doi:10.1126/science.aar4120.
42. Hertel, R.; Rodríguez, D.P.; Hollensteiner, J.; Dietrich, S.; Leimbach, A.; Hoppert, M.; Liesegang, H.; Volland, S. Genome-Based Identification of Active Prophage Regions by Next Generation Sequencing in *Bacillus licheniformis* DSM13. *PLOS ONE* **2015**, *10*, e0120759. doi:10.1371/journal.pone.0120759.
43. Matos, R.C.; Lapaque, N.; Rigottier-Gois, L.; Debarbieux, L.; Meylheuc, T.; Gonzalez-Zorn, B.; Repoila, F.; Lopes, M.d.F.; Serror, P. *Enterococcus faecalis* Prophage Dynamics and Contributions to Pathogenic Traits. *PLOS Genetics* **2013**, *9*, 1–16. doi:10.1371/journal.pgen.1003539.
44. de Maagd, R.A.; Bravo, A.; Crickmore, N. How *Bacillus thuringiensis* has evolved specific toxins to colonize the insect world. *TRENDS in Genetics* **2001**, *17*, 193–199.
45. Haft, D.H.; DiCuccio, M.; Badretdin, A.; Brover, V.; Chetvernin, V.; O'Neill, K.; Li, W.; Chitsaz, F.; Derbyshire, M.K.; Gonzales, N.R.; others. RefSeq: an update on prokaryotic genome annotation and curation. *Nucleic acids research* **2018**, *46*, D851–D860.
46. Larkin, M.A.; Blackshields, G.; Brown, N.P.; Chenna, R.; Mcgettigan, P.A.; McWilliam, H.; Valentin, F.; Wallace, I.M.; Wilm, A.; Lopez, R.; Thompson, J.D.; Gibson, T.J.; Higgins, D.G. Clustal W and Clustal X version 2.0. *Bioinformatics* **2007**, *23*, 2947–2948. doi:10.1093/bioinformatics/btm404.
47. Kumar, S.; Stecher, G.; Li, M.; Knyaz, C.; Tamura, K. MEGA X: Molecular evolutionary genetics analysis across computing platforms. *Molecular Biology and Evolution* **2018**, *35*, 1547–1549. doi:10.1093/molbev/msy096.
48. Stothard, P. The sequence manipulation suite: JavaScript programs for analyzing and formatting protein and DNA sequences. *BioTechniques* **2000**, *28*. doi:10.2144/00286ir01.

49. Gabler, F.; Nam, S.Z.; Till, S.; Mirdita, M.; Steinegger, M.; Söding, J.; Lupas, A.N.; Alva, V. Protein Sequence Analysis Using the MPI Bioinformatics Toolkit. *Current Protocols in Bioinformatics* **2020**, *72*, e108. doi:10.1002/cpbi.108.
50. Siguier, P.; Perochon, J.; Lestrade, L.; Mahillon, J.; Chandler, M. ISfinder: the reference centre for bacterial insertion sequences. *Nucleic acids research* **2006**, *34*, 32–36. doi:10.1093/nar/gkj014.

6.2 Talk: IDOPS, a profile HMM-based tool to detect pesticidal sequences

A talk titled "IDOPS, a profile HMM-based tool to detect pesticidal sequences" was presented at the Bioinformatics: from algorithms to applications (BIATA) virtual conference 2021. The video is available at: https://www.youtube.com/watch?v=DynABY_Qn5U

6.3 Poster

The following graphical abstract and poster were presented at the VAAM Annual Conference 2022.

GEORG-AUGUST-UNIVERSITÄT
GÖTTINGEN

Genomic and Applied Microbiology

Poster ID: eP308

Download as interactive powerpoint show

Comparative Genomics of Chromosomally-Encoded Pesticidal Genes Reveals a Novel Prophage-Associated *cry* Cassette

Anat Lev Hacohen[1*], Stefani Díaz Valerio[1], Jacqueline Hollensteiner[1], Raphael Schöppe[1*] and Heiko Liesegang[1]

[1] Genomic and Applied Microbiology & Göttingen Genomics Laboratory, Institute of Microbiology and Genetics, Georg-August University of Göttingen, Göttingen, Germany
*Research conducted during studies at [1]

Stefani Díaz Valerio
sdiazva@gwdg.de

Click me!

THE MICRO CHRONICLE

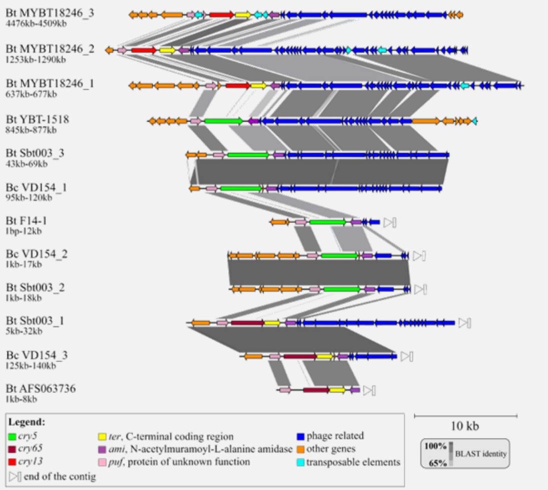
VAAM 2022
Corona Edition

Comparative Genomics of Chromosomally-Encoded Pesticidal Genes Reveals a Novel Prophage-Associated *cry* Cassette

IT'S IN THE CHROMOSOME!

Bacillus thuringiensis **under investigation**

Initial report confirms 12 cassette occurrences



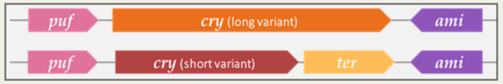
FEATURED THE CASSETTE ELEMENTS

puf: gene for protein of unknown function.

cry: coding Cry5, Cry13 (nematocidal) and Cry65 (unknown target) toxins. Either as long variant or as split version with adjacent C-terminal domains (*ter*).

Conserved surrounding but diverse toxins

ami: N-acetylmuramoyl-L-alanine amidase.



UNEXPECTED!

Suspicious *Siphoviridae*-like prophage regions were found in proximity to the cassette. Do they play a role in transmission?

EDITORIAL Can we infect BT strains with a toxin carrying virus?

Exclusive 22.02.2022
16:15-17:45
Poster Session 8

Poster ID: eP308

GEORG-AUGUST-UNIVERSITÄT
GÖTTINGEN

Comparative Genomics of Chromosomally-Encoded Pesticidal Genes Reveals a Novel Prophage-Associated *cry* Cassette

Genomic and Applied Microbiology

1. State of the art
Toxins from *Bacillus thuringiensis* (BT)

- Highly specific against invertebrate pests
- Nontoxic to humans
- Eco-friendly
- Most successful biopesticide worldwide
- Used to control vectors of human diseases
- Usually encoded on conjugative plasmids

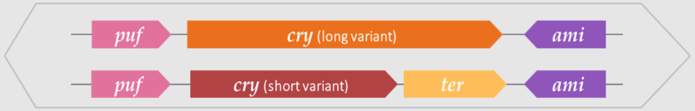
Only few found on chromosomal replicons

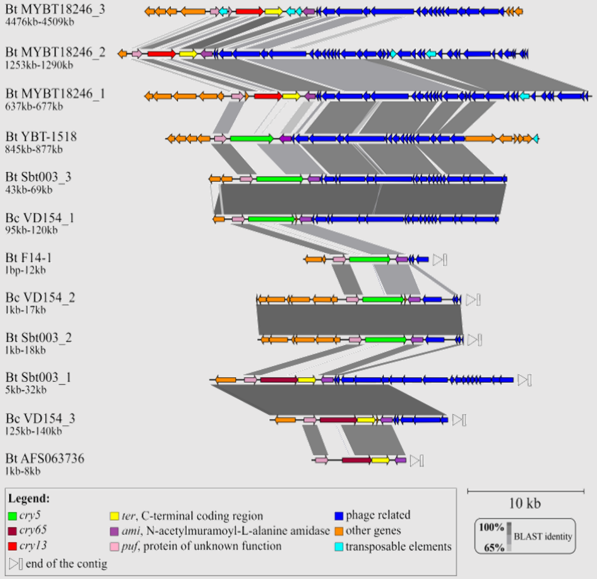
GOAL: Identify them and characterize their genetic environment

2. Methods
Comparative genomics of *cry* genes loci

Can we...

Find more occurrences of chromosomal *cry* genes?	Collection of *B. thuringiensis* genomes IDOPS NCBI BLAST *IDentification Of Pesticidal Sequences*
Identify conserved co-located elements around them?	
Characterize the toxin gene and unveil evolutionary traits?	•CDVist •Isfinder •ClustalW2 •Ident and Sim •Phaster •ProphageHunter





Legend:

🟩 *cry5*
🟥 *cry65*
🟥 *cry13*
▷▯ end of the contig
🟨 *ter*, C-terminal coding region
🟪 *ami*, N-acetylmuramoyl-L-alanine amidase
■ *puf*, protein of unknown function
🟦 phage related
🟧 other genes
■ transposable elements

3. Results
Conserved surrounding but diverse toxins

puf
- Protein of unknown function
- HTH domain identified

cry
- Cry5, Cry13 (both nematocidal) and Cry65 (unknown target)
- Either long variant or as split version with adjacent C-terminal domain (*ter*)

ami
- N-acetylmuramoyl-L-alanine amidase
- Also reported on plasmid cassettes

12 cassette instances across BT genomes
Adjacent to ***Siphoviridae*-like** prophage regions

4. Conclusions and Outlook
Chromosomally encoded *cry* genes

We found...

- A genetic cassette of conserved accessory elements around dynamic *cry* genes → *Roles of puf and ami genes?*
- Evidence of recombination resulting in diverse 3-domain Cry toxins → *Evolution dynamics behind cry diversification?*
- Cassette-associated *Siphoviridae*-like prophage regions → *Is the prophage involved in transmission?*

How can we use these insights to improve biopesticidal BT strains?

Click me!

Anat Lev Hacohen[1], Stefani Díaz Valerio[1], Jacqueline Hollensteiner[1], Raphael Schöppe[1] and Heiko Liesegang[1]

[1] Genomic and Applied Microbiology & Göttingen Genomics Laboratory, Institute of Microbiology and Genetics, Georg-August University of Göttingen, Göttingen, Germany

Stefani Díaz Valerio
sdiazva@gwdg.de

Appendix A

General Microbiology and Molecular Biology methods

A.1 Bacterial Strains

Strain	Description	Source/Reference
B. subtilis 168	Laboratory type strain, Δ*trpC2*	AG-Daniel Strain Collection
B. subtilis Δ6	Genome-reduced strain	AG-Daniel Strain Collection
B. subtilis 168 ΔyqfD	Deletion of gene *yqfD*	This work
E. coli S17-1		AG-Daniel Strain Collection
B. licheniformis MW3 ΔyqfD	Deletion of genes *yqfD*, *hsdR1*, *hsdR2*	AG-Daniel Strain Collection
B. licheniformis DSM13	Type strain	AG-Daniel Strain Collection
B. pumilus DSM27	Type strain	AG-Daniel Strain Collection
B. pumilus MS32		AB Enzymes GmbH, Darmstadt, Germany

A.2 Media and Solutions

LB media/plates	For 500 ml
Trypton	5 g
Yeast Extract	2.5 g
NaCl	2.5 g
*Agar	7 g

*Add in order to prepare agar plates

#416	For 500 ml
Tryptone	10 g
Yeast extract	5 g
NaCl	5 g
*2M Sucrose solution	50 ml

* Autoclave separately and add it afterwards.

Expression Mix	10,5 ml
Yeast Extract 5%	5 ml
Casamino-acids 10%	2.5 ml
Tryptophan 5mg/ml	500 µL
Sterile Water	2.5ml

2M Sucrose solution	For 500 ml
Sucrose	342 g

2× SMM buffer	For 500 ml at 2X
Sucrose	171.5 g
Maleic acid	3.35 g
MgCl2 6H2O	4.06 g

Adjust pH to 6,4 with NaOH

AB3/Penassay Broth	For 500 ml at 4X
Peptone	10 g
Yeast Extract	3 g
Beef Extract	3 g
Dextrose	2 g
NaCl	7 g
K2HPO4	7.35 g
KH2PO4	2.6 5g

PEG solution	For 100 ml
PEG 6000	40 g
2X SMM	50 ml

DM3 Plates	For 1L, autoclave separately
Agar	8 g in 200 ml H2O
Sodium succinate	91g in 500 ml H2O, pH 7.3
Casamino acids	5 g in 100 ml H2O
Yeast Extract	5 g in 50 ml H2O
K2HPO4 and KH2PO4	3,5 g and 1,5 g in 100 ml H2O
Glucose	6 g in 30 ml H2O
MgCl2	1.9 g in 20 ml H2O

After autoclave, add 5 ml of filter-sterilized BSA 2%

BSA solution	For 50 ml
BSA	2.5 g
2X SMM	50 ml

Glycerol recovery media	For 500ml
Tryptone	5 g
Yeast extract	2.5 g
NaCl	2.5 g
Sorbitol	45.5 g
Mannitol	34.5 g
Glycerol	50 g
Agar	7.5 g

A.3 Clean-up of enzymatic reactions

MagSi-NGS $^{\text{PREP}}$ Plus

MagSi beads were used to purify PCR products according to suppliers instructions as provided at `https://www.magtivio.com/magsi-ngsprep-plus/`, with the minor modification of doing the final elution step with pre-warmed water (50°C-60°C)

QIAquick Gel Extraction Kit

Whenever PCR reactions produced unspecific products, the band corresponding to the desired fragment was excised from the gel and purified with the QIAquick Gel Extraction Kit following suppliers instructions.

A.4 Oligos

Primers used to generate a deletion cassette for the *yqfD* gene of *B. pumilus* DSM27.

ID	Sequence	Length (bp)
H1_forward	gtgaagaattagTGCGATATTCGTAAGGAGAAGAAAATT	39
H1_reverse	AATATCGCActaattcttcacacttctcccctcc	34
H2_forward	CCTGTCAACtgaggagactagagaatgacagaacatttac	40
H2_reverse	gtctcctcaGTTGACAGGGACATCTGAATCC	31
Control_F	gagatgtccttgatctctgcaagc	24
Control_R	agccagctcctcagcgtaag	20
FlanK_F	cgcacagaacgatgaaacggc	21
Flank_R	catgaacggcattgacgactgc	22
Nested_F	acctgatggagatttcttgcagt	23
Nested_R	ctgttccagctggtccaattcca	23
Nested_R2	ggccaatggtttttaccctgatt	23

Primers used to generate a deletion cassette for the *yqfD* gene of *B. subtilis* 168.

ID	Sequence 5′-3′	Length (bp)
SDV0004	gtcattcttcGTTGACAGGGACATCTGAATCCCTC	35
SDV0005	gttgtgaaaaatTGCGATATTCGTAAGGAGAAGAAAATTC	40
SDV0006	TCCCTGTCAACgaagaatgacagaacatttacttgcg	37
SDV0007	atcataaaaatgacccgataacgc	24
SDV0008	CGAATATCGCAatttttcacaacatttccccctcgg	36
SDV0009	tcaagaacaggaaatgcgtgcc	22
SDV0010	gcaaagcaaacataatggcag	21
SDV0011	CACGATGACTCAGTATGTAATGC	23
SDV0012	CGAAGTCTCATTGAGTGTGTC	21
SDV0013	gatattgcagatgtagatatcggc	24
SDV0014	ctggtctagctcaatcatagaaat	24
SDV0015	gaatattggcgtcatggattacat	24

Appendix B

Bioinformatic commands

This work used the Scientific Compute Cluster at GWDG, the joint data center of Max Planck Society for the Advancement of Science (MPG) and University of Göttingen.

B.1 Comparative Genomics

B.1.1 PyANI

Version 0.2.11

```
>average_nucleotide_identity.py -o output_ANIm -i input_genomes
--workers 16 --labels labels.txt -g -f
```

B.1.2 PGAP

Version 2021-07-01.build5508

```
>python3.7 pgap.py GENOME.fasta --no-internet -D /singularity
-c 24 Genome_input.yaml
```

B.1.3 ProteinOrtho

Version 6.0.31

```
>proteinortho -project=PROJECT\_NAME -cpus=16 -singles /*.faa
```

B.1.4 antiSMASH

Version 6.0.1

```
>antismash --cpus 24 --fullhmmer --tigrfam --cb-general
--cb-knownclusters --pfam2go --asf --cb-subclusters --smcog-trees
--cc-mibig --output-dir /OUT GENOME.gbk
```

B.1.5 CRISPRCasFinder

Version 4.2.20

```
>singularity exec -B \$PWD crisprcasfinder.sif perl \CPF
-so \SO -cf \CF -drpt \DRPT -rpts \RPTS -cas -def G
-out output -in GENOME.fna
```

B.1.6 ISEScan

Version 1.7.2.3

```
>isescan.py --removeShortIS --seqfile GENOME.fna --output OUT
--nthread 16
```

B.1.7 PhiSpy

Version 4.2.19 with hmmer version 3.3.2
First press VOGs database:

```
>hmmpress vogs210.hmm
>phispy -o OUT -p NAME --threads 16 --phmms vogs210.hmm
--color --output_choice 11 GENOME.gbk
```

B.1.8 RGI

Version rgi-5.2.1+CARD-3.1.4

```
>rgi load -i card.json --local
>rgi main --input_sequence GENOME.faa --output_file OUT
--input_type protein --clean --local --alignment_tool DIAMOND
--num_threads 16 --split_prodigal_jobs
```

B.1.9 REBASE

BLAST version 2.2.31+ and REBASE version as in 02.2022

```
>makeblastdb -in rebase_prot.fasta -dbtype prot
>blastp -num_threads 12 -evalue 1e-25 -out ID_rebase
-db rebase_prot.fasta -query ID.faa -outfmt '6 qseqid sseqid
length qlen slen mismatch evalue  score bitscore qcovs
qcovhsp stitle' -max_target_seqs 1
```

B.1.10 MEROPS

Blast version 2.2.31+ and MEROPS 12.1

```
>blastp -num_threads 12 -evalue 1e-25 -out OUT.txt
-db merops_scanlib.fasta -query ID.faa -outfmt '6 qseqid
sseqid length qlen slen mismatch evalue score bitscore qcovs
qcovhsp stitle' -max_target_seqs 1
```

B.1.11 COGClassifier

Version 1.2.0

```
>COGclassifier -i ID.faa -o ID/ -d /COGClassifier/ -t 12
```

B.1.12 KOfamScan

Version 1.3.0

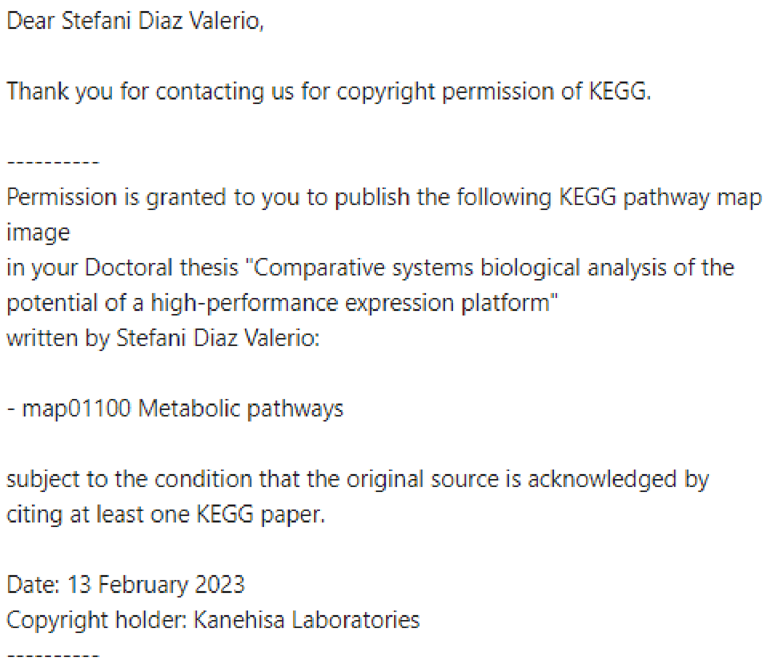
```
>exec_annotation --cpu 24 --tmp-dir $TMP_LOCAL -k KL -p PROF
-f mapper -o ID.txt ID.fa
```

B.1.13 SignalP

Version 6.0

```
>signalp6 -ff ID.faa -od ID/ -m fast -org other -fmt txt -wp 12 -bs 12
```

B.2 KEGG Copyright Permission



Dear Stefani Diaz Valerio,

Thank you for contacting us for copyright permission of KEGG.

Permission is granted to you to publish the following KEGG pathway map
image
in your Doctoral thesis "Comparative systems biological analysis of the
potential of a high-performance expression platform"
written by Stefani Diaz Valerio:

- map01100 Metabolic pathways

subject to the condition that the original source is acknowledged by
citing at least one KEGG paper.

Date: 13 February 2023
Copyright holder: Kanehisa Laboratories

FIGURE B.1: Copyright permission by KEGG.

B.3 RNAseq

Here "ID" stands for the identification of the corresponding organism of the analysis, for example MW3 (for *B. licheniformis* MW3). "LIBRARIES" stands for the RNA-seq libraries used as input. "CONDITIONS" in the READemption command stands for the sampling time points.

B.3.1 FastP

Version 0.20.1

```
>fastp -i ID_R1.fastq.gz -o ID_p1.fastq -I ID_R2.fastq.gz
-O ID_p2.fastq --unpaired1 ID_unpaired.fastq --unpaired2
ID_unpaired.fastq --failed_out ID_failed.fastq
--detect_adapter_for_pe -p -w 16 -L -c -h ID_fastp.html
-j ID_fastp.json
```

B.3.2 RSEQC

Version 4.0.0

```
>tin.py -i /alignments/ -r ./GENOME.bed
```

B.3.3 SortmeRNA

Version 4.3.3

```
>sortmerna -ref ID.fasta -reads ID_fastq.gz --threads 24 --fastx
--workdir /ID/ --aligned ID_rRNA --other ID_nr
```

B.3.4 READemption

Version 1.0.10

```
>reademption align --processes 12 --paired_end --fastq --reverse_complement
-f /ProjectID
>reademption coverage --processes 12 -f /ProjectID
>reademption gene_quanti --add_antisense --processes 12 -f /ProjectID
>reademption deseq -l "LIBRARIES" -c "CONDITIONS" -f /ProjectID
>reademption viz_align --paired_end -f /ProjectID
>reademption viz_gene_quanti --paired_end -f /ProjectID
>reademption viz_deseq -f /ProjectID
```

B.3.5 Annogesic

Version 1.0.22

```
>annogesic transcript --project_path /path/ --annotation_files ID.gff
--modify_transcript merge_overlap --frag_libs LIBRARIES
--replicate_frag all_3 --compare_feature_genome gene CDS
>annogesic terminator --project_path /path/ --annotation_files ID.gff
--transcript_files ID_transcript.gff --fasta_files ID.fa
--frag_libs LIBRARIES --replicate_frag all_3
>annogesic sorf --project_path /path/ --annotation_files ID.gff
--transcript_files ID_transcript.gff  --fasta_files ID.fa
--frag_libs LIBRARIES --replicate_frag all_3
>annogesic srna --project_path /path/ --filter_info sec_str blast_srna
--compute_sec_structures --srna_database_path /RNA.fa  --srna_format
--sorf_files ID_sORF.gff --annotation_files ID.gff --transcript_files
ID_transcript.gff --fasta_files ID.fa --terminator_files ID_term.gff
--frag_libs LIBRARIES --replicate_frag all_3
>annogesic srna_target --project_path /path/ --program RNAplex
RNAup IntaRNA --top 50 --parallels_rnaplex 24
--parallels_rnaup 24 --parallels_intarna 24 --annotation_files ID.gff
--fasta_files ID.fa  --srna_files ID_sRNA.gff
```

B.3.6 RFAM

RFAM database version 14.7 and Infernal 1.1.4

```
>cmscan --cut_ga --rfam --nohmmonly --fmt 2 --cpu 32 --clanin CLAN -Z IDZ
--tblout tblout.txt -o out.txt RFAM ID.fna
```

B.3.7 DP_GP_cluster

Version v.0.1

```
>DP_GP_cluster.py -i TPM_zasinh_mean.txt -o OUTPUT -p pdf
--true_times --plot -n 1000 --fast --cluster_uncertainty_estimate
```

Bibliography

[1] Teppei Abe et al. "tmRNA-dependent trans-translation is required for sporulation in Bacillus subtilis". In: *Molecular microbiology* 69.6 (2008), pp. 1491–1498.

[2] EFSA Panel on Additives et al. "Efficacy of Bacillus subtilis DSM 28343 as a zootechnical additive (gut flora stabiliser) for calves for rearing". In: *EFSA Journal* 17.7 (2019), e05793.

[3] Yvonne Agersø et al. "Putative antibiotic resistance genes present in extant Bacillus licheniformis and Bacillus paralicheniformis strains are probably intrinsic and part of the ancient resistome". In: *PloS one* 14.1 (2019), e0210363.

[4] Sajia Akhter, Ramy K Aziz, and Robert A Edwards. "PhiSpy: a novel algorithm for finding prophages in bacterial genomes that combines similarity- and composition-based strategies". In: *Nucleic acids research* 40.16 (2012), e126–e126.

[5] Sandra S Albrecht et al. "Lessons Learned From Dear Pandemic, a Social Media–Based Science Communication Project Targeting the COVID-19 Infodemic". In: *Public Health Reports* 137.3 (2022), pp. 449–456.

[6] Brian P Alcock et al. "CARD 2020: antibiotic resistome surveillance with the comprehensive antibiotic resistance database". In: *Nucleic acids research* 48.D1 (2020), pp. D517–D525.

[7] Svetlana Alexeeva et al. "Genomics of tailless bacteriophages in a complex lactic acid bacteria starter culture". In: *International Dairy Journal* 114 (2021), p. 104900.

[8] Omer S Alkhnbashi et al. "CRISPR-Cas bioinformatics". In: *Methods* 172 (2020), pp. 3–11.

[9] Stephen F Altschul et al. "Basic local alignment search tool". In: *Journal of molecular biology* 215.3 (1990), pp. 403–410.

[10] Miguel A Andrade et al. "NEAT: a domain duplicated in genes near the components of a putative Fe3+ siderophore transporter from Gram-positive pathogenic bacteria". In: *Genome biology* 3.9 (2002), pp. 1–5.

[11] Gabriele Aquilina et al. "Scientific Opinion on the safety and efficacy of Bacillus subtilis PB6 (Bacillus subtilis) as a feed additive for chickens for fattening". In: (2009).

[12] David R Arahal. "Whole-genome analyses: average nucleotide identity". In: *Methods in microbiology*. Vol. 41. Elsevier, 2014, pp. 103–122.

[13] Takuya Aramaki et al. "KofamKOALA: KEGG ortholog assignment based on profile HMM and adaptive score threshold". In: *Bioinformatics* 36.7 (2020), pp. 2251–2252.

[14] Paul G Arnison et al. "Ribosomally synthesized and post-translationally modified peptide natural products: overview and recommendations for a universal nomenclature". In: *Natural product reports* 30.1 (2013), pp. 108–160.

[15] Trond Erik Vee Aune and Finn Lillelund Aachmann. "Methodologies to increase the transformation efficiencies and the range of bacteria that can be transformed". In: *Applied microbiology and biotechnology* 85.5 (2010), pp. 1301–1313.

[16] Ramy K Aziz et al. "The RAST Server: rapid annotations using subsystems technology". In: *BMC genomics* 9.1 (2008), pp. 1–15.

[17] Juri Niño Bach and Marc Bramkamp. "Flotillins functionally organize the bacterial membrane". In: *Molecular microbiology* 88.6 (2013), pp. 1205–1217.

[18] Rolf Backofen and Wolfgang R Hess. "Computational prediction of sRNAs and their targets in bacteria". In: *RNA biology* 7.1 (2010), pp. 33–42.

[19] Florian Baier and Nobuhiko Tokuriki. "Connectivity between catalytic landscapes of the metallo-β-lactamase superfamily". In: *Journal of Molecular Biology* 426.13 (2014), pp. 2442–2456.

[20] Giulia Barbieri et al. "Interplay of CodY and ScoC in the regulation of major extracellular protease genes of Bacillus subtilis". In: *Journal of bacteriology* 198.6 (2016), pp. 907–920.

[21] Francisco Fábio Cavalcante Barros et al. "Production of enzymes from agroindustrial wastes by biosurfactant-producing strains of Bacillus subtilis". In: *Biotechnology research international* 2013 (2013).

[22] Christian J Barton and Mark A Merolli. "It is time to replace publish or perish with get visible or vanish: opportunities where digital and social media can reshape knowledge translation". In: *British Journal of Sports Medicine* 53.10 (2019), pp. 594–598.

[23] Alex Bateman et al. "The Pfam protein families database". In: *Nucleic acids research* 32.suppl_1 (2004), pp. D138–D141.

[24] Jay J Van Bavel et al. "Using social and behavioural science to support COVID-19 pandemic response". In: *Nature human behaviour* 4.5 (2020), pp. 460–471.

[25] David H Bechhofer and Murray P Deutscher. "Bacterial ribonucleases and their roles in RNA metabolism". In: *Critical reviews in biochemistry and molecular biology* 54.3 (2019), pp. 242–300.

[26] Joel G Belasco. "All things must pass: contrasts and commonalities in eukaryotic and bacterial mRNA decay". In: *Nature reviews Molecular cell biology* 11.7 (2010), pp. 467–478.

[27] Martin Benda et al. "Influence of the ABC transporter YtrBCDEF of Bacillus subtilis on competence, biofilm formation and cell wall thickness". In: *Frontiers in microbiology* 12 (2021), p. 587035.

[28] Stephen D Bentley and Julian Parkhill. "Comparative genomic structure of prokaryotes". In: *Annual review of genetics* 38.1 (2004), pp. 771–791.

[29] Kevin Blighe, Sharmila Rana, and Myles Lewis. "EnhancedVolcano: Publication-ready volcano plots with enhanced colouring and labeling". In: *R package version* 1.0 (2019).

[30] Kai Blin et al. "antiSMASH 6.0". In: *Nucleic Acids Research* (2021).

[31] Matthias Blum et al. "The InterPro protein families and domains database: 20 years on". In: *Nucleic acids research* 49.D1 (2021), pp. D344–D354.

[32] Mirjam Boonstra et al. "Analyses of competent and non-competent subpopulations of Bacillus subtilis reveal yhfW, yhxC and ncRNAs as novel players in competence". In: *Environmental microbiology* 22.6 (2020), pp. 2312–2328.

[33] Claudia Borgmeier, Johannes Bongaerts, and Friedhelm Meinhardt. "Genetic analysis of the Bacillus licheniformis degSU operon and the impact of regulatory mutations on protease production". In: *Journal of biotechnology* 159.1-2 (2012), pp. 12–20.

[34] Marc Bramkamp and Daniel Lopez. "Exploring the existence of lipid rafts in bacteria". In: *Microbiology and Molecular Biology Reviews* 79.1 (2015), pp. 81–100.

[35] Sabine Brantl and Peter Müller. "Cis-and trans-encoded small regulatory RNAs in Bacillus subtilis". In: *Microorganisms* 9.9 (2021), p. 1865.

[36] Shaun R Brinsmade et al. "Hierarchical expression of genes controlled by the Bacillus subtilis global regulatory protein CodY". In: *Proceedings of the National Academy of Sciences* 111.22 (2014), pp. 8227–8232.

[37] Ulf Brockmeier et al. "Systematic screening of all signal peptides from Bacillus subtilis: a powerful strategy in optimizing heterologous protein secretion in Gram-positive bacteria". In: *Journal of molecular biology* 362.3 (2006), pp. 393–402.

[38] Jeremy S Brown and David W Holden. "Iron acquisition by Gram-positive bacterial pathogens". In: *Microbes and infection* 4.11 (2002), pp. 1149–1156.

[39] Boyke Bunk et al. "Bacillus megaterium and other bacilli: industrial applications". In: *Encyclopedia of industrial biotechnology: bioprocess, bioseparation, and cell technology* (2009), pp. 1–15.

[40] Terry W Burns, D John O'Connor, and Susan M Stocklmayer. "Science communication: a contemporary definition". In: *Public understanding of science* 12.2 (2003), pp. 183–202.

[41] Anke Busch, Andreas S Richter, and Rolf Backofen. "IntaRNA: efficient prediction of bacterial sRNA targets incorporating target site accessibility and seed regions". In: *Bioinformatics* 24.24 (2008), pp. 2849–2856.

[42] Dongbo Cai et al. "A novel strategy to improve protein secretion via overexpression of the SppA signal peptide peptidase in Bacillus licheniformis". In: *Microbial cell factories* 16.1 (2017), pp. 1–10.

[43] Robert Caldwell et al. *Correlation between Bacillus subtilis scoC phenotype and gene expression determined using microarrays for transcriptome analysis*. 2001.

[44] Christiam Camacho et al. "BLAST+: architecture and applications". In: *BMC bioinformatics* 10.1 (2009), pp. 1–9.

[45] Rosa Capita and Carlos Alonso-Calleja. "Antibiotic-resistant bacteria: a challenge for the food industry". In: *Critical reviews in food science and nutrition* 53.1 (2013), pp. 11–48.

[46] Sherwood Casjens. "Prophages and bacterial genomics: what have we learned so far?" In: *Molecular microbiology* 49.2 (2003), pp. 277–300.

[47] Fidel Antonio Castro-Smirnov et al. "Physical interactions between DNA and sepiolite nanofibers, and potential application for DNA transfer into mammalian cells". In: *Scientific reports* 6.1 (2016), pp. 1–14.

[48] Amy T Cavanagh and Karen M Wassarman. "6S RNA, a global regulator of transcription in Escherichia coli, Bacillus subtilis, and beyond". In: *Annual review of microbiology* 68 (2014), pp. 45–60.

[49] Shing Chang and Stanley N Cohen. "High frequency transformation of Bacillus subtilis protoplasts by plasmid DNA". In: *Molecular and General Genetics MGG* 168.1 (1979), pp. 111–115.

[50] Aleksandra Checinska, Malcolm Burbank, and Andrzej J Paszczynski. "Protection of Bacillus pumilus spores by catalases". In: *Applied and Environmental Microbiology* 78.18 (2012), pp. 6413–6422.

[51] James Chen et al. "6S RNA mimics B-form DNA to regulate Escherichia coli RNA polymerase". In: *Molecular cell* 68.2 (2017), pp. 388–397.

[52] Shifu Chen et al. "fastp: an ultra-fast all-in-one FASTQ preprocessor". In: *Bioinformatics* 34.17 (2018), pp. i884–i890.

[53] Jae Woong Choi et al. "Enhanced production of recombinant proteins with Corynebacterium glutamicum by deletion of insertion sequences (IS elements)". In: *Microbial Cell Factories* 14.1 (2015), pp. 1–12.

[54] Ambrose R Cole et al. "Clostridium perfringens ε-toxin shows structural similarity to the pore-forming toxin aerolysin". In: *Nature structural & molecular biology* 11.8 (2004), pp. 797–798.

[55] European Commission, Directorate-General for Research, and Innovation. *A sustainable bioeconomy for Europe : strengthening the connection between economy, society and the environment : updated bioeconomy strategy*. Publications Office, 2018. DOI: doi/10.2777/792130.

[56] Claire Concannon and Muriel Grenon. *Researchers: share your passion for science!* 2016.

[57] Ciaran Condon et al. "Comparison of the expression of the seven ribosomal RNA operons in Escherichia coli." In: *The EMBO journal* 11.11 (1992), pp. 4175–4185.

[58] Gene Ontology Consortium. "The gene ontology resource: 20 years and still GOing strong". In: *Nucleic acids research* 47.D1 (2019), pp. D330–D338.

[59] UniProt Consortium. "Reorganizing the protein space at the Universal Protein Resource (UniProt)". In: *Nucleic acids research* 40.D1 (2012), pp. D71–D75.

[60] Jessika Consuegra et al. "Insertion-sequence-mediated mutations both promote and constrain evolvability during a long-term experiment with bacteria". In: *Nature communications* 12.1 (2021), pp. 1–12.

[61] Fabiano Jares Contesini, Ricardo Rodrigues de Melo, and Hélia Harumi Sato. "An overview of Bacillus proteases: from production to application". In: *Critical reviews in biotechnology* 38.3 (2018), pp. 321–334.

[62] Tyrrell Conway et al. "Unprecedented high-resolution view of bacterial operon architecture revealed by RNA sequencing". In: *MBio* 5.4 (2014), e01442–14.

[63] Hervé Corvellec, Alison F Stowell, and Nils Johansson. "Critiques of the circular economy". In: *Journal of Industrial Ecology* 26.2 (2022), pp. 421–432.

[64] David Couvin et al. "CRISPRCasFinder, an update of CRISRFinder, includes a portable version, enhanced performance and integrates search for Cas proteins". In: *Nucleic acids research* 46.W1 (2018), W246–W251.

[65] Neil Crickmore et al. “A structure-based nomenclature for Bacillus thuringiensis and other bacteria-derived pesticidal proteins”. In: *Journal of invertebrate pathology* 186 (2021), p. 107438.

[66] Mary A Crum, B Trevor Sewell, and Michael J Benedik. “Bacillus pumilus cyanide dihydratase mutants with higher catalytic activity”. In: *Frontiers in Microbiology* 7 (2016), p. 1264.

[67] Wenjing Cui et al. “Exploitation of Bacillus subtilis as a robust workhorse for production of heterologous proteins and beyond”. In: *World Journal of Microbiology and Biotechnology* 34.10 (2018), pp. 1–19.

[68] Peter H Culviner, Chantal K Guegler, and Michael T Laub. “A simple, cost-effective, and robust method for rRNA depletion in RNA-sequencing studies”. In: *MBio* 11.2 (2020), e00010–20.

[69] Iuliia Danilova and Margarita Sharipova. “The practical potential of Bacilli and their enzymes for industrial production”. In: *Frontiers in microbiology* 11 (2020), p. 1782.

[70] IV Danilova et al. “Optimization of Electroporation Conditions for Bacillus pumilus 3–19 Strain”. In: *BioNanoScience* (2022), pp. 1–5.

[71] Aaron CE Darling et al. “Mauve: multiple alignment of conserved genomic sequence with rearrangements”. In: *Genome research* 14.7 (2004), pp. 1394–1403.

[72] Christian Degering et al. “Optimization of protease secretion in Bacillus subtilis and Bacillus licheniformis by screening of homologous and heterologous signal peptides”. In: *Applied and environmental microbiology* 76.19 (2010), pp. 6370–6376.

[73] Arnold L Demain and Aiqi Fang. “The natural functions of secondary metabolites”. In: *History of modern biotechnology I* (2000), pp. 1–39.

[74] Yi Deng et al. “Improved inducible expression of Bacillus naganoensis pullulanase from recombinant Bacillus subtilis by enhancer regulation”. In: *Protein Expression and Purification* 148 (2018), pp. 9–15.

[75] Stefani Díaz-Valerio et al. “IDOPS, a profile HMM-based tool to detect pesticidal sequences and compare their genetic context”. In: *Frontiers in Microbiology* 12 (2021).

[76] Jan Maarten van Dijl et al. “Functional genomic analysis of the Bacillus subtilis Tat pathway for protein secretion”. In: *Journal of biotechnology* 98.2-3 (2002), pp. 243–254.

[77] JanMaarten van Dijl and Michael Hecker. *Bacillus subtilis: from soil bacterium to super-secreting cell factory*. 2013.

[78] David Dubnau. “Genetic competence in Bacillus subtilis”. In: *Microbiological reviews* 55.3 (1991), pp. 395–424.

[79] Christopher A Dunlap et al. “Bacillus velezensis is not a later heterotypic synonym of Bacillus amyloliquefaciens; Bacillus methylotrophicus, Bacillus amyloliquefaciens subsp. plantarum and ‘Bacillus oryzicola’are later heterotypic synonyms of Bacillus velezensis based on phylogenomics”. In: *International journal of systematic and evolutionary microbiology* 66.3 (2016), pp. 1212–1217.

[80] Sylvain Durand et al. “Identification of an RNA sponge that controls the RoxS riboregulator of central metabolism in Bacillus subtilis”. In: *Nucleic acids research* 49.11 (2021), pp. 6399–6419.

[81] Ilhan Cem Duru et al. “RNA editing, RNA modifications, and transcriptional units in Listeria monocytogenes”. In: (2022).

[82] H Auguste Dutcher and Rahul Raghavan. “Origin, evolution, and loss of bacterial small RNAs”. In: *Microbiology spectrum* 6.2 (2018), pp. 6–2.

[83] Michael B Eisen et al. “Cluster analysis and display of genome-wide expression patterns”. In: *Proceedings of the National Academy of Sciences* 95.25 (1998), pp. 14863–14868.

[84] Francesca Ermoli et al. “SwrA as global modulator of the two-component system DegSU in Bacillus subtilis”. In: *Research in Microbiology* 172.6 (2021), p. 103877.

[85] Martín Espariz et al. “Taxonomic identity resolution of highly phylogenetically related strains and selection of phylogenetic markers by using genome-scale methods: the Bacillus pumilus group case”. In: *PLoS One* 11.9 (2016), e0163098.

[86] Alicia Fajardo and José L Martínez. “Antibiotics as signals that trigger specific bacterial responses”. In: *Current opinion in microbiology* 11.2 (2008), pp. 161–167.

[87] Anton Farr. “Multi-species comparison of sRNA related regulation in selected representatives of the *Bacillus* genus”. MA thesis. Georg-August-University Göttingen, "2022".

[88] Anton Robert Georg Farr. *Nextflow pipeline for RNAseq based multi-species comparison of sRNA regulation*. Version 0.1.1. Nov. 2022. DOI: 10.5281/zenodo.7384159. URL: https://doi.org/10.5281/zenodo.7384159.

[89] Annaleigh Ohrt Fehler et al. “Flagella disruption in Bacillus subtilis increases amylase production yield”. In: *Microbial Cell Factories* 21.1 (2022), pp. 1–11.

[90] Y Feng et al. “Fermentation of starch for enhanced alkaline protease production by constructing an alkalophilic Bacillus pumilus strain”. In: *Applied microbiology and biotechnology* 57.1 (2001), pp. 153–160.

[91] Andrea Feucht, Louise Evans, and Jeff Errington. “Identification of sporulation genes by genome-wide analysis of the σE regulon of Bacillus subtilis”. In: *Microbiology* 149.10 (2003), pp. 3023–3034.

[92] Sebastien Fleurier et al. “rRNA operon multiplicity as a bacterial genome stability insurance policy”. In: *Nucleic Acids Research* (2022).

[93] Konrad U Foerstner et al. “Environments shape the nucleotide composition of genomes”. In: *EMBO reports* 6.12 (2005), pp. 1208–1213.

[94] Enzymes EFSA Panel on Food Contact Materials et al. “Safety evaluation of the food enzyme α-amylase from the genetically modified Bacillus licheniformis strain DP-Dzb52”. In: *EFSA Journal* 19.4 (2021), e06564.

[95] Enzymes EFSA Panel on Food Contact Materials et al. “Safety evaluation of the food enzyme α-amylase from the genetically modified Bacillus licheniformis strain NZYM-BC”. In: *EFSA Journal* 20.7 (2022), e07370.

[96] Enzymes EFSA Panel on Food Contact Materials et al. “Safety evaluation of the food enzyme β-galactosidase from the genetically modified Bacillus licheniformis strain NZYM-BT”. In: *EFSA Journal* 20.7 (2022), e07358.

[97] Enzymes EFSA Panel on Food Contact Materials et al. “Safety evaluation of the food enzyme β-galactosidase from the genetically modified Bacillus licheniformis strain NZYM-BT”. In: *EFSA Journal* 20.7 (2022), e07358.

[98] Enzymes EFSA Panel on Food Contact Materials et al. "Safety evaluation of the food enzyme maltogenic α-amylase from the genetically modified Bacillus subtilis strain ROM". In: *EFSA Journal* 19.6 (2021), e06634.

[99] Flavourings EFSA Panel on Food Contact Materials Enzymes et al. "Safety evaluation of the food enzyme pullulanase from genetically modified Bacillus subtilis strain NZYM-AK". In: *EFSA Journal* 15.8 (2017), e04895.

[100] Flavourings EFSA Panel on Food Contact Materials Enzymes et al. "Safety evaluation of the food enzyme xylanase from a genetically modified Bacillus subtilis strain TD 160 (229)". In: *EFSA Journal* 16.1 (2018), e05008.

[101] Processing Aids (CEP) EFSA Panel on Food Contact Materials Enzymes et al. "Safety evaluation of the food enzyme alpha-amylase from a genetically modified Bacillus subtilis (strain NBA)". In: *EFSA Journal* 17.5 (2019), e05681.

[102] Processing Aids (CEP) EFSA Panel on Food Contact Materials Enzymes et al. "Safety evaluation of the food enzyme endo-1, 3 (4)-β-glucanase from the genetically modified Bacillus subtilis strain DP-Ezm28". In: *EFSA Journal* 19.3 (2021), e06431.

[103] Samuel C Forster et al. "RNA-eXpress annotates novel transcript features in RNA-seq data". In: *Bioinformatics* 29.6 (2013), pp. 810–812.

[104] Konrad U Förstner, Jörg Vogel, and Cynthia M Sharma. "READemption—a tool for the computational analysis of deep-sequencing–based transcriptome data". In: *Bioinformatics* 30.23 (2014), pp. 3421–3423.

[105] Jéssyca Freitas-Silva et al. "Peeling the Layers Away: The Genomic Characterization of Bacillus pumilus 64-1, an Isolate With Antimicrobial Activity From the Marine Sponge Plakina cyanorosea (Porifera, Homoscleromorpha)". In: *Frontiers in microbiology* (2021), p. 3402.

[106] Roland Freudl. "Signal peptides for recombinant protein secretion in bacterial expression systems". In: *Microbial cell factories* 17.1 (2018), pp. 1–10.

[107] Cecilie From, Victor Hormazabal, and Per Einar Granum. "Food poisoning associated with pumilacidin-producing Bacillus pumilus in rice". In: *International journal of food microbiology* 115.3 (2007), pp. 319–324.

[108] Xiaoteng Fu et al. "Bacillus pumilus Group Comparative Genomics: Toward Pangenome Features, Diversity, and Marine Environmental Adaptation". In: *Frontiers in microbiology* 12 (2021), p. 1084.

[109] Toni Gabaldón and Eugene V Koonin. "Functional and evolutionary implications of gene orthology". In: *Nature Reviews Genetics* 14.5 (2013), pp. 360–366.

[110] Michael Y Galperin, Roger Higdon, and Eugene Kolker. "Interplay of heritage and habitat in the distribution of bacterial signal transduction systems". In: *Molecular BioSystems* 6.4 (2010), pp. 721–728.

[111] Michael Y Galperin et al. "Microbial genome analysis: the COG approach". In: *Briefings in bioinformatics* 20.4 (2019), pp. 1063–1070.

[112] Paul P Gardner et al. "Rfam: Wikipedia, clans and the "decimal" release". In: *Nucleic acids research* 39.suppl_1 (2010), pp. D141–D145.

[113] Adrian Sven Geissler et al. "CRISPRi screen for enhancing heterologous α-amylase yield in Bacillus subtilis". In: *BioRxiv* (2022).

[114] Jens Georg and Wolfgang R Hess. "Widespread antisense transcription in prokaryotes". In: *Regulating with RNA in Bacteria and Archaea* (2018), pp. 191–210.

[115] Jens Georg et al. "The power of cooperation: Experimental and computational approaches in the functional characterization of bacterial sRNAs". In: *Molecular Microbiology* 113.3 (2020), pp. 603–612.

[116] Matthias Gimpel and Sabine Brantl. "Dual-function sRNA encoded peptide SR1P modulates moonlighting activity of B. subtilis GapA". In: *RNA biology* 13.9 (2016), pp. 916–926.

[117] Uri Gophna et al. "No evidence of inhibition of horizontal gene transfer by CRISPR–Cas on evolutionary timescales". In: *The ISME journal* 9.9 (2015), pp. 2021–2027.

[118] Johan Goris et al. "DNA–DNA hybridization values and their relationship to whole-genome sequence similarities". In: *International journal of systematic and evolutionary microbiology* 57.1 (2007), pp. 81–91.

[119] Susan Gottesman and Gisela Storz. "Bacterial small RNA regulators: versatile roles and rapidly evolving variations". In: *Cold Spring Harbor perspectives in biology* 3.12 (2011), a003798.

[120] Jonathan Gottschall. *The storytelling animal: How stories make us human*. Houghton Mifflin Harcourt, 2012.

[121] Sam Griffiths-Jones et al. "Rfam: an RNA family database". In: *Nucleic acids research* 31.1 (2003), pp. 439–441.

[122] Anastasiia Grigoreva et al. "Identification and characterization of andalusicin: N-terminally dimethylated class III lantibiotic from Bacillus thuringiensis sv. andalousiensis". In: *Iscience* 24.5 (2021), p. 102480.

[123] Masja Nierop Groot, Frank Nieboer, and Tjakko Abee. "Enhanced transformation efficiency of recalcitrant Bacillus cereus and Bacillus weihenstephanensis isolates upon in vitro methylation of plasmid DNA". In: *Applied and environmental microbiology* 74.24 (2008), pp. 7817–7820.

[124] Kirk J Grubbs et al. "Large-scale bioinformatics analysis of Bacillus genomes uncovers conserved roles of natural products in bacterial physiology". In: *MSystems* 2.6 (2017), e00040–17.

[125] Yang Gu et al. "Advances and prospects of Bacillus subtilis cellular factories: from rational design to industrial applications". In: *Metabolic engineering* 50 (2018), pp. 109–121.

[126] Jan Gundlach et al. "Control of potassium homeostasis is an essential function of the second messenger cyclic di-AMP in Bacillus subtilis". In: *Science signaling* 10.475 (2017), eaal3011.

[127] Hermann Hämmerle et al. "Impact of Hfq on the Bacillus subtilis transcriptome". In: *PloS one* 9.6 (2014), e98661.

[128] Lin-Li Han et al. "Disruption of the pleiotropic gene scoC causes transcriptomic and phenotypical changes in Bacillus pumilus BA06". In: *BMC genomics* 20.1 (2019), pp. 1–11.

[129] Lin-Li Han et al. "Transcriptome profiling analysis reveals metabolic changes across various growth phases in Bacillus pumilus BA06". In: *BMC microbiology* 17.1 (2017), pp. 1–13.

[130] Stefan Handtke et al. "Bacillus pumilus reveals a remarkably high resistance to hydrogen peroxide provoked oxidative stress". In: *PLoS One* 9.1 (2014), e85625.

[131] Colin R Harwood and Rocky Cranenburgh. "Bacillus protein secretion: an unfolding story". In: *Trends in microbiology* 16.2 (2008), pp. 73–79.

[132] Colin R Harwood and Yoshimi Kikuchi. "The ins and outs of Bacillus proteases: activities, functions and commercial significance". In: *FEMS Microbiology Reviews* 46.1 (2022), fuab046.

[133] Graham F Hatfull and Roger W Hendrix. "Bacteriophages and their genomes". In: *Current opinion in virology* 1.4 (2011), pp. 298–303.

[134] Hasmik Hayrapetyan et al. "Comparative genomics of iron-transporting systems in Bacillus cereus strains and impact of iron sources on growth and biofilm formation". In: *Frontiers in Microbiology* 7 (2016), p. 842.

[135] Julian D Hegemann and Roderich D Süssmuth. "Matters of class: coming of age of class III and IV lanthipeptides". In: *RSC Chemical Biology* 1.3 (2020), pp. 110–127.

[136] Nadja Heidrich, Isabella Moll, and Sabine Brantl. "In vitro analysis of the interaction between the small RNA SR1 and its primary target ahrC mRNA". In: *Nucleic acids research* 35.13 (2007), pp. 4331–4346.

[137] Gunnar von Heijne. "Life and death of a signal peptide". In: *Nature* 396.6707 (1998), pp. 111–113.

[138] R Hertel, S Volland, and H Liesegang. "Conjugative reporter system for the use in Bacillus licheniformis and closely related Bacilli". In: *Letters in applied microbiology* 60.2 (2015), pp. 162–167.

[139] Robert Hertel et al. "Small RNA mediated repression of subtilisin production in Bacillus licheniformis". In: *Scientific reports* 7.1 (2017), pp. 1–11.

[140] Kerstin Hoffmann et al. "Facilitation of direct conditional knockout of essential genes in Bacillus licheniformis DSM13 by comparative genetic analysis and manipulation of genetic competence". In: *Applied and environmental microbiology* 76 (2010), pp. 5046–5057.

[141] RK Holmes and MG Jobling. "Genetics: conjugation". In: *Barons medical microbiology* 4 (1996).

[142] Erin S Honsa, Anthony W Maresso, and Sarah K Highlander. "Molecular and evolutionary analysis of NEAr-iron Transporter (NEAT) domains". In: *PLoS One* 9.8 (2014), e104794.

[143] Jens Hör, Stanislaw A Gorski, and Jörg Vogel. "Bacterial RNA biology on a genome scale". In: *Molecular cell* 70.5 (2018), pp. 785–799.

[144] Robert M Horton et al. "Engineering hybrid genes without the use of restriction enzymes: gene splicing by overlap extension". In: *Gene* 77.1 (1989), pp. 61–68.

[145] Xue-Jia Hu et al. "Prokaryotic and highly-repetitive WD40 proteins: A systematic study". In: *Scientific reports* 7.1 (2017), pp. 1–13.

[146] Franziska Huff et al. "The restriction modification system of Bacillus licheniformis MS1 and generation of a readily transformable deletion mutant". In: *Applied microbiology and biotechnology* 101.21 (2017), pp. 7933–7944.

[147] Annelore Huyghe, Mirjam Knockaert, and Martin Obschonka. “Unraveling the “passion orchestra” in academia”. In: *Journal of Business Venturing* 31.3 (2016), pp. 344–364.

[148] Hanne-Leena Hyyryläinen, Matti Sarvas, and Vesa P Kontinen. “Transcriptome analysis of the secretion stress response of Bacillus subtilis”. In: *Applied microbiology and biotechnology* 67 (2005), pp. 389–396.

[149] Sajid Iqbal, John Vollmers, and Hussnain Ahmed Janjua. “Genome mining and comparative genome analysis revealed niche-specific genome expansion in antibacterial bacillus pumilus strain SF-4”. In: *Genes* 12.7 (2021), p. 1060.

[150] Tarequl Islam et al. “Biosynthesis, Molecular Regulation, and Application of Bacilysin Produced by Bacillus Species”. In: *Metabolites* 12.5 (2022), p. 397.

[151] Mareike Jakobs et al. “Unravelling the genetic basis for competence development of auxotrophic Bacillus licheniformis 9945A strains”. In: *Microbiology* 160.10 (2014), pp. 2136–2147.

[152] Lina Jakutyte-Giraitiene and Giedrius Gasiunas. “Design of a CRISPR-Cas system to increase resistance of Bacillus subtilis to bacteriophage SPP1”. In: *Journal of Industrial Microbiology and Biotechnology* 43.8 (2016), pp. 1183–1188.

[153] Brian D Janssen and Christopher S Hayes. “The tmRNA ribosome-rescue system”. In: *Advances in protein chemistry and structural biology* 86 (2012), pp. 151–191.

[154] Albert Jeltsch. “Maintenance of species identity and controlling speciation of bacteria: a new function for restriction/modification systems?” In: *Gene* 317 (2003), pp. 13–16.

[155] Min Jiang, Roberto Grau, and Marta Perego. “Differential processing of propeptide inhibitors of Rap phosphatases in Bacillus subtilis”. In: *Journal of bacteriology* 182.2 (2000), pp. 303–310.

[156] Kayla A Johnson and Arjun Krishnan. “Robust normalization and transformation techniques for constructing gene coexpression networks from RNA-seq data”. In: *Genome biology* 23.1 (2022), pp. 1–26.

[157] Keith A Jolley et al. “Ribosomal multilocus sequence typing: universal characterization of bacteria from domain to strain”. In: *Microbiology* 158.Pt 4 (2012), p. 1005.

[158] Philip Jones et al. “InterProScan 5: genome-scale protein function classification”. In: *Bioinformatics* 30.9 (2014), pp. 1236–1240.

[159] Ioanna Kalvari et al. “Rfam 14: expanded coverage of metagenomic, viral and microRNA families”. In: *Nucleic Acids Research* 49.D1 (2021), pp. D192–D200.

[160] Kazimierz Kaminski and Teresa Sokolowska. “The probable identity of bacilysin and tetaine”. In: *The Journal of Antibiotics* 26.3 (1973), pp. 184–185.

[161] Kyoko Kanamaru, Sophie Stephenson, and Marta Perego. “Overexpression of the PepF oligopeptidase inhibits sporulation initiation in Bacillus subtilis”. In: *Journal of bacteriology* 184.1 (2002), pp. 43–50.

[162] Shigehiko Kanaya et al. “Studies of codon usage and tRNA genes of 18 unicellular organisms and quantification of Bacillus subtilis tRNAs: gene expression level and species-specific diversity of codon usage based on multivariate analysis”. In: *Gene* 238.1 (1999), pp. 143–155.

[163] Minoru Kanehisa and Susumu Goto. "KEGG: kyoto encyclopedia of genes and genomes". In: *Nucleic acids research* 28.1 (2000), pp. 27–30.

[164] Minoru Kanehisa, Yoko Sato, and Kanae Morishima. "BlastKOALA and GhostKOALA: KEGG tools for functional characterization of genome and metagenome sequences". In: *Journal of molecular biology* 428.4 (2016), pp. 726–731.

[165] Usha Kantiwal and Janmejay Pandey. "Efficient Inhibition of Bacterial Biofilm Through Interference of Protein–Protein Interaction of Master Regulator Proteins: a Proof of Concept Study with SinR-SinI Complex of Bacillus subtilis". In: *Applied Biochemistry and Biotechnology* (2022), pp. 1–21.

[166] Klemens Kappel and Sebastian Jon Holmen. "Why science communication, and does it work? A taxonomy of science communication aims and a survey of the empirical evidence". In: *Frontiers in Communication* 4 (2019), p. 55.

[167] Peter D Karp et al. "The metacyc database". In: *Nucleic acids research* 30.1 (2002), pp. 59–61.

[168] Kenneth C Keiler and Nitya S Ramadoss. "Bifunctional transfer-messenger RNA". In: *Biochimie* 93.11 (2011), pp. 1993–1997.

[169] Michael J Kempf et al. "Recurrent isolation of hydrogen peroxide-resistant spores of Bacillus pumilus from a spacecraft assembly facility". In: *Astrobiology* 5.3 (2005), pp. 391–405.

[170] Eleanor Hadley Kershaw et al. "The sustainable path to a circular bioeconomy". In: *Trends in Biotechnology* 39.6 (2021), pp. 542–545.

[171] R Khaneja et al. "Carotenoids found in Bacillus". In: *Journal of applied microbiology* 108.6 (2010), pp. 1889–1902.

[172] Keitarou Kimura, Lam-Son Phan Tran, and Kazumi Funane. "Loss of poly-γ-glutamic Acid Synthesis of Bacillus subtilis (natto) Due to IS4Bsu1 Translocation to swrA Gene". In: *Food Science and Technology Research* 17.5 (2011), pp. 447–451.

[173] Julian Kirchherr, Denise Reike, and Marko Hekkert. "Conceptualizing the circular economy: An analysis of 114 definitions". In: *Resources, conservation and recycling* 127 (2017), pp. 221–232.

[174] Joel A Klappenbach, John M Dunbar, and Thomas M Schmidt. "rRNA operon copy number reflects ecological strategies of bacteria". In: *Applied and environmental microbiology* 66.4 (2000), pp. 1328–1333.

[175] Gary L Kleman and William R Strohl. "High cell density and high-productivity microbial fermentation". In: *Current Opinion in Biotechnology* 3.2 (1992), pp. 93–98.

[176] Ichizo Kobayashi. "Behavior of restriction–modification systems as selfish mobile elements and their impact on genome evolution". In: *Nucleic acids research* 29.18 (2001), pp. 3742–3756.

[177] Kazuo Kobayashi. "SlrR/SlrA controls the initiation of biofilm formation in Bacillus subtilis". In: *Molecular microbiology* 69.6 (2008), pp. 1399–1410.

[178] Ahasanul Kobir et al. "Phosphorylation of B acillus subtilis gene regulator AbrB modulates its DNA-binding properties". In: *Molecular Microbiology* 92.5 (2014), pp. 1129–1141.

[179] Marc AB Kolkman and Eugenio Ferrari. "The fate of extracellular proteins tagged by the SsrA system of Bacillus subtilis". In: *Microbiology* 150.2 (2004), pp. 427–436.

[180] Eugene V Koonin. “Orthologs, paralogs, and evolutionary genomics”. In: *Annu. Rev. Genet.* 39 (2005), pp. 309–338.

[181] Evguenia Kopylova, Laurent Noé, and Hélène Touzet. “SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data”. In: *Bioinformatics* 28.24 (2012), pp. 3211–3217.

[182] Sanna Koskiniemi et al. “Selection-driven gene loss in bacteria”. In: *PLoS genetics* 8.6 (2012), e1002787.

[183] Ákos T Kovács. “Bacillus subtilis”. In: *Trends in Microbiology* 27.8 (2019), pp. 724–725.

[184] Tobias Küppers et al. “Developing a new production host from a blueprint: Bacillus pumilus as an industrial enzyme producer”. In: *Microbial cell factories* 13.1 (2014), pp. 1–11.

[185] Jangyul Kwak, Hao Jiang, and Kathleen E Kendrick. “Transformation using in vivo and in vitro methylation in Streptomyces griseus”. In: *FEMS microbiology letters* 209.2 (2002), pp. 243–248.

[186] Mckee L. “One Type of Soil Bacteria Performs Two Important Jobs to Help Us Produce Healthy Food”. In: *Front. Young Minds* 8:554161 (2020). DOI: 10.3389/frym.2020.554161.

[187] Sebastian Laass et al. “Characterization of the transcriptome of Haloferax volcanii, grown under four different conditions, with mixed RNA-Seq”. In: *PLoS One* 14.4 (2019), e0215986.

[188] Marcus Lechner et al. “Proteinortho: detection of (co-) orthologs in large-scale analysis”. In: *BMC bioinformatics* 12.1 (2011), pp. 1–9.

[189] Joshua Lederberg and Edward L Tatum. “Gene recombination in Escherichia coli”. In: *Nature* 158.4016 (1946), p. 558.

[190] Martin Lehnik-Habrink et al. “DEAD-Box RNA helicases in Bacillus subtilis have multiple functions and act independently from each other”. In: *Journal of bacteriology* 195.3 (2013), pp. 534–544.

[191] Simon Leonard et al. “APERO: a genome-wide approach for identifying bacterial small RNAs from RNA-Seq data”. In: *Nucleic acids research* 47.15 (2019), e88–e88.

[192] Corinne Lévi-Meyrueis, Katalin Fodor, and Pierre Schaeffer. “Polyethyleneglycol-induced transformation of Bacillus subtilis protoplasts by bacterial chromosomal DNA”. In: *Molecular and General Genetics MGG* 179.3 (1980), pp. 589–594.

[193] Wuju Li et al. “Predicting sRNAs and their targets in bacteria”. In: *Genomics, proteomics & bioinformatics* 10.5 (2012), pp. 276–284.

[194] Lars Lilge et al. “Expression of degQ gene and its effect on lipopeptide production as well as formation of secretory proteases in Bacillus subtilis strains”. In: *MicrobiologyOpen* 10.5 (2021), e1241.

[195] Ta-Hui Lin, Shih-Chien Huang, and Gwo-Chyuan Shaw. “Reexamining transcriptional regulation of the Bacillus subtilis htpX gene and the ykrK gene, encoding a novel type of transcriptional regulator, and redefining the YkrK operator”. In: *Journal of bacteriology* 194.24 (2012), pp. 6758–6765.

[196] Hong Liu et al. “Bacillus pumilus LZP02 promotes rice root growth by improving carbohydrate metabolism and phenylpropanoid biosynthesis”. In: *Molecular Plant-Microbe Interactions* 33.10 (2020), pp. 1222–1231.

[197] Long Liu et al. "How to achieve high-level expression of microbial enzymes: strategies and perspectives". In: *Bioengineered* 4.4 (2013), pp. 212–223.

[198] Yang Liu et al. "Phylogenetic diversity of the Bacillus pumilus group and the marine ecotype revealed by multilocus sequence analysis". In: *PloS one* 8.11 (2013), e80097.

[199] Yong-Cheng Liu et al. "Characterization of a protease hyper-productive mutant of Bacillus pumilus by comparative genomic and transcriptomic analysis". In: *Current Microbiology* 77.11 (2020), pp. 3612–3622.

[200] Zhongzhong Liu et al. "The highly modified microcin peptide plantazolicin is associated with nematicidal activity of Bacillus amyloliquefaciens FZB42". In: *Applied microbiology and biotechnology* 97.23 (2013), pp. 10081–10090.

[201] Jonathan Livny and Matthew K Waldor. "Identification of small RNAs in diverse bacterial species". In: *Current opinion in microbiology* 10.2 (2007), pp. 96–101.

[202] NA Logan and RCW Berkeley. "Identification of Bacillus strains using the API system". In: *Microbiology* 130.7 (1984), pp. 1871–1882.

[203] Yvonne Lokko et al. "Biotechnology and the bioeconomy—Towards inclusive and sustainable industrial development". In: *New Biotechnology* 40 (2018). Bioeconomy, pp. 5–10. ISSN: 1871-6784. DOI: `https://doi.org/10.1016/j.nbt.2017.06.005`. URL: `https://www.sciencedirect.com/science/article/pii/S1871678416326206`.

[204] Daniel López and Roberto Kolter. "Extracellular signals that define distinct and coexisting cell fates in Bacillus subtilis". In: *FEMS microbiology reviews* 34.2 (2010), pp. 134–149.

[205] EH Madslien et al. "Lichenysin is produced by most Bacillus licheniformis strains". In: *Journal of applied microbiology* 115.4 (2013), pp. 1068–1080.

[206] Jacques Mahillon and Michael Chandler. "Insertion sequences". In: *Microbiology and molecular biology reviews* 62.3 (1998), pp. 725–774.

[207] AU Mahmood, J Greenman, and AH Scragg. "Orange and potato peel extracts: Analysis and use as Bacillus substrates for the production of extracellular enzymes in continuous culture". In: *Enzyme and microbial technology* 22.2 (1998), pp. 130–137.

[208] Martin CJ Maiden et al. "Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms". In: *Proceedings of the National Academy of Sciences* 95.6 (1998), pp. 3140–3145.

[209] Berenike Maier. "Competence and transformation in Bacillus subtilis". In: *Current Issues in Molecular Biology* 37.1 (2020), pp. 57–76.

[210] Kira S Makarova et al. "An updated evolutionary classification of CRISPR–Cas systems". In: *Nature Reviews Microbiology* 13.11 (2015), pp. 722–736.

[211] Prashant Mali, Kevin M Esvelt, and George M Church. "Cas9 as a versatile tool for engineering biology". In: *Nature methods* 10.10 (2013), pp. 957–963.

[212] Scott Mann and Yi-Ping Phoebe Chen. "Bacterial genomic G+ C composition-eliciting environmental adaptation". In: *Genomics* 95.1 (2010), pp. 7–15.

[213] Nicola Manzo et al. "Pigmentation and sporulation are alternative cell fates in Bacillus pumilus SF214". In: *PLoS One* 8.4 (2013), e62093.

[214] Carmelita Nora Marbaniang and Jörg Vogel. "Emerging roles of RNA modifications in bacteria". In: *Current opinion in microbiology* 30 (2016), pp. 50–57.

[215] Ruben AT Mars et al. "Small regulatory RNA-induced growth rate heterogeneity of Bacillus subtilis". In: *PLoS genetics* 11.3 (2015), e1005046.

[216] David J Martínez-Cano et al. "Evolution of small prokaryotic genomes". In: *Frontiers in microbiology* 5 (2015), p. 742.

[217] S Marzi and P Romby. "RNA mimicry, a decoy for regulatory proteins". In: *Molecular microbiology* 83.1 (2012), pp. 1–6.

[218] Andres Maser et al. "Amino acids are key substrates to Escherichia coli BW25113 for achieving high specific growth rate". In: *Research in microbiology* 171.5-6 (2020), pp. 185–193.

[219] Andrew G McArthur et al. "The comprehensive antibiotic resistance database". In: *Antimicrobial agents and chemotherapy* 57.7 (2013), pp. 3348–3357.

[220] Ian C McDowell et al. "Clustering gene expression time series data using an infinite Gaussian process mixture model". In: *PLoS computational biology* 14.1 (2018), e1005896.

[221] Marnix H Medema and Michael A Fischbach. "Computational approaches to natural product discovery". In: *Nature chemical biology* 11.9 (2015), pp. 639–648.

[222] Wilfried JJ Meijer et al. "Multiple layered control of the conjugation process of the Bacillus subtilis plasmid pLS20". In: *Frontiers in Molecular Biosciences* 8 (2021).

[223] Flávia Mandolesi Pereira de Melo et al. "Antifungal compound produced by the cassava endophyte Bacillus pumilus MAIIIM4A". In: *Scientia Agricola* 66 (2009), pp. 583–592.

[224] Marcus Miethke et al. "Iron starvation triggers the stringent response and induces amino acid biosynthesis for bacillibactin production in Bacillus subtilis". In: *Journal of bacteriology* 188.24 (2006), pp. 8655–8657.

[225] Morgan E Milton et al. "The solution structures and interaction of SinR and SinI: elucidating the mechanism of action of the master regulator switch for biofilm formation in Bacillus subtilis". In: *Journal of molecular biology* 432.2 (2020), pp. 343–357.

[226] Ralf Moeller et al. "Role of pigmentation in protecting Bacillus sp. endospores against environmental UV radiation". In: *FEMS microbiology ecology* 51.2 (2005), pp. 231–236.

[227] Katie J Molohon et al. "Plantazolicin is an ultranarrow-spectrum antibiotic that targets the Bacillus anthracis membrane". In: *ACS infectious diseases* 2.3 (2015), pp. 207–220.

[228] Katie J Molohon et al. "Structure determination and interception of biosynthetic intermediates for the plantazolicin class of highly discriminating antibiotics". In: *ACS chemical biology* 6.12 (2011), pp. 1307–1313.

[229] Kyung Moon et al. "Identification of BvgA-Dependent and BvgA-Independent Small RNAs (sRNAs) in Bordetella pertussis Using the Prokaryotic sRNA Prediction Toolkit ANNOgesic". In: *Microbiology Spectrum* 9.2 (2021), e00044–21.

[230] Yuki Moriya et al. “KAAS: an automatic genome annotation and pathway reconstruction server”. In: *Nucleic acids research* 35.suppl_2 (2007), W182–W185.

[231] Sampriti Mukherjee and Daniel B Kearns. “The structure and regulation of flagella in Bacillus subtilis”. In: *Annual review of genetics* 48 (2014), p. 319.

[232] Peter Müller et al. “A new role for CsrA: promotion of complex formation between an sRNA and its mRNA target in Bacillus subtilis”. In: *RNA biology* 16.7 (2019), pp. 972–987.

[233] Pascal Mülner et al. “Profiling for bioactive peptides and volatiles of plant growth promoting strains of the Bacillus subtilis complex of industrial relevance”. In: *Frontiers in Microbiology* (2020), p. 1432.

[234] Hye-won Na et al. “Structural and biochemical analyses of the metallo-β-lactamase fold protein YhfI from Bacillus subtilis”. In: *Biochemical and Biophysical Research Communications* 519.1 (2019), pp. 35–40.

[235] Toshiro Nagai et al. “A new IS 4 family insertion sequence, IS 4Bsu 1, responsible for genetic instability of poly-γ-glutamic acid production in Bacillus subtilis”. In: *Journal of Bacteriology* 182.9 (2000), pp. 2387–2392.

[236] Hannes Nahrstedt et al. “Strain development in Bacillus licheniformis: construction of biologically contained mutants deficient in sporulation and DNA repair”. In: *Journal of biotechnology* 119.3 (2005), pp. 245–254.

[237] Catherine Nannan et al. “Bacilysin within the Bacillus subtilis group: gene prevalence versus antagonistic activity against Gram-negative foodborne pathogens”. In: *Journal of Biotechnology* 327 (2021), pp. 28–35.

[238] Eric P Nawrocki, Diana L Kolbe, and Sean R Eddy. “Infernal 1.0: inference of RNA alignments”. In: *Bioinformatics* 25.10 (2009), pp. 1335–1337.

[239] Bruno TL Nichio, Jeroniza Nunes Marchaukoski, and Roberto Tadeu Raittz. “New tools in orthology analysis: a brief review of promising perspectives”. In: *Frontiers in genetics* 8 (2017), p. 165.

[240] Reindert Nijland and Oscar P Kuipers. “Optimization of protein secretion by Bacillus subtilis”. In: *Recent Patents on Biotechnology* 2.2 (2008), pp. 79–87.

[241] Mor Nitzan, Rotem Rehani, and Hanah Margalit. “Integration of bacterial small RNAs in regulatory networks”. In: *Annual review of biophysics* 46 (2017), pp. 131–148.

[242] Mitsuo Ogura, Hirofumi Yoshikawa, and Taku Chibazakura. “Regulation of the response regulator gene degU through the binding of SinR/SlrR and exclusion of SinR/SlrR by DegU in Bacillus subtilis”. In: *Journal of bacteriology* 196.4 (2014), pp. 873–881.

[243] Pedro H Oliveira, Marie Touchon, and Eduardo PC Rocha. “Regulation of genetic flux between bacteria by restriction–modification systems”. In: *Proceedings of the National Academy of Sciences* 113.20 (2016), pp. 5658–5663.

[244] Gülay Özcengiz and İsmail Öğülür. “Biochemistry, genetics and regulation of bacilysin biosynthesis and its significance more than an antibiotic”. In: *New biotechnology* 32.6 (2015), pp. 612–619.

[245] Adrien Pain et al. “An assessment of bacterial small RNA target prediction programs”. In: *RNA biology* 12.5 (2015), pp. 509–513.

[246] Sarah E Palmer and Renato A Schibeci. “What conceptions of science communication are espoused by science research funding bodies?” In: *Public Understanding of Science* 23.5 (2014), pp. 511–527.

[247] Tracy Palmer and Ben C Berks. “The twin-arginine translocation (Tat) protein export pathway”. In: *Nature Reviews Microbiology* 10.7 (2012), pp. 483–496.

[248] Kai Papenfort and Jörg Vogel. “Multiple target regulation by small noncoding RNAs rewires gene expression at the post-transcriptional level”. In: *Research in microbiology* 160.4 (2009), pp. 278–287.

[249] J Parrado et al. “Proteomic analysis of enzyme production by Bacillus licheniformis using different feather wastes as the sole fermentation media”. In: *Enzyme and Microbial Technology* 57 (2014), pp. 1–7.

[250] Tiago Pedreira, Christoph Elfmann, and Jörg Stülke. “The current state of Subti Wiki, the database for the model organism Bacillus subtilis”. In: *Nucleic Acids Research* 50.D1 (2022), pp. D875–D882.

[251] Alessandro Pellis et al. “Evolving biocatalysis to meet bioeconomy challenges and opportunities”. In: *New biotechnology* 40 (2018), pp. 154–169.

[252] Mengxue Peng and Zhihong Liang. “Degeneration of industrial bacteria caused by genetic instability”. In: *World Journal of Microbiology and Biotechnology* 36.8 (2020), pp. 1–16.

[253] Qin Peng, Yihui Yuan, and Meiying Gao. “Bacillus pumilus, a novel ginger rhizome rot pathogen in China”. In: *Plant Disease* 97.10 (2013), pp. 1308–1315.

[254] Thomas Nordahl Petersen et al. “SignalP 4.0: discriminating signal peptides from transmembrane regions”. In: *Nature methods* 8.10 (2011), pp. 785–786.

[255] Olga E Petrova et al. “Comparative evaluation of rRNA depletion procedures for the improved analysis of bacterial biofilm and mixed pathogen culture transcriptomes”. In: *Scientific reports* 7.1 (2017), pp. 1–15.

[256] Andrew M Phillips et al. “Molecular and cell biological analysis of SwrB in Bacillus subtilis”. In: *Journal of bacteriology* 203.17 (2021), e00227–21.

[257] Hualiang Pi and John D Helmann. “Sequential induction of Fur-regulated genes in response to iron limitation in Bacillus subtilis”. In: *Proceedings of the National Academy of Sciences* 114.48 (2017), pp. 12785–12790.

[258] René van der Ploeg et al. “High-salinity growth conditions promote Tat-independent secretion of Tat substrates in Bacillus subtilis”. In: *Applied and environmental microbiology* 78.21 (2012), pp. 7733–7744.

[259] Zoltán Prágai and Colin R Harwood. “Regulatory interactions between the Pho and σB-dependent general stress regulons of Bacillus subtilis”. In: *Microbiology* 148.5 (2002), pp. 1593–1602.

[260] Heike Preis et al. “CodY activates transcription of a small RNA in Bacillus subtilis”. In: *Journal of bacteriology* 191.17 (2009), pp. 5446–5457.

[261] Leighton Pritchard et al. “Genomics and taxonomy in diagnostics for food security: soft-rotting enterobacterial plant pathogens”. In: *Analytical Methods* 8.1 (2016), pp. 12–24.

[262] Yimin Qiu et al. “Improvement of lichenysin production in Bacillus licheniformis by replacement of native promoter of lichenysin biosynthesis operon and medium optimization”. In: *Applied microbiology and biotechnology* 98 (2014), pp. 8895–8903.

[263] Michael Rachinger et al. “Size unlimited markerless deletions by a transconjugative plasmid-system in Bacillus licheniformis”. In: *Journal of biotechnology* 167.4 (2013), pp. 365–369.

[264] Sandrine Ragu, Olivier Piétrement, and Bernard S Lopez. "Binding of DNA to Natural Sepiolite: Applications in Biotechnology and Perspectives". In: *Clays and Clay Minerals* 69.5 (2021), pp. 633–640.

[265] Adriana Ravagnani, Christopher L Finan, and Michael Young. "A novel firmicute protein family related to the actinobacterial resuscitation-promoting factors by non-orthologous domain displacement". In: *BMC genomics* 6.1 (2005), pp. 1–14.

[266] Neil D Rawlings and Alex Bateman. "How to use the MEROPS database and website to help understand peptidase specificity". In: *Protein Science* 30.1 (2021), pp. 83–92.

[267] Neil D Rawlings et al. "The MEROPS database of proteolytic enzymes, their substrates and inhibitors in 2017 and a comparison with peptidases in the PANTHER database". In: *Nucleic acids research* 46.D1 (2018), pp. D624–D632.

[268] Aiming Ren and Dinshaw J Patel. "c-di-AMP binds the ydaO riboswitch in two pseudo-symmetry–related pockets". In: *Nature chemical biology* 10.9 (2014), pp. 780–786.

[269] Jun Ren, Dokyun Na, and Seung Min Yoo. "Optimization of chemico-physical transformation methods for various bacterial species using diverse chemical compounds and nanomaterials". In: *Journal of biotechnology* 288 (2018), pp. 55–60.

[270] Jun Ren et al. "Combined chemical and physical transformation method with RbCl and sepiolite for the transformation of various bacterial species". In: *Journal of microbiological methods* 135 (2017), pp. 48–51.

[271] Michael W Rey et al. "Complete genome sequence of the industrial bacterium Bacillus licheniformis and comparisons with closely related Bacillusspecies". In: *Genome biology* 5.10 (2004), pp. 1–12.

[272] Paul J Riesenman and Wayne L Nicholson. "Role of the spore coat layers in Bacillus subtilis spore resistance to hydrogen peroxide, artificial UV-C, UV-B, and solar UV radiation". In: *Applied and environmental microbiology* 66.2 (2000), pp. 620–626.

[273] Binetti del Rio et al. "Multiplex PCR for the detection and identification of dairy bacteriophages in milk". In: *Food microbiology* 24.1 (2007), pp. 75–81.

[274] Richard J Roberts and Dana Macelis. "REBASE—restriction enzymes and methylases". In: *Nucleic Acids Research* 26.1 (1998), pp. 338–350.

[275] Richard J Roberts et al. "REBASE—a database for DNA restriction and modification: enzymes, genes and genomes". In: *Nucleic acids research* 43.D1 (2015), pp. D298–D299.

[276] Eduardo PC Rocha and David Bikard. "Microbial defenses against mobile genetic elements and viruses: Who defends whom from what?" In: *PLoS biology* 20.1 (2022), e3001514.

[277] BRK Roller, SF Stoddard, and TM Schmidt. *Exploiting rRNA operon copy number to investigate bacterial reproductive strategies. Nat Microbiol 1: 16160*. 2016.

[278] Diego Romero et al. "Antibiotics as signal molecules". In: *Chemical reviews* 111.9 (2011), pp. 5492–5505.

[279] Carlos A Ruiz-Perez, Roth E Conrad, and Konstantinos T Konstantinidis. "MicrobeAnnotator: a user-friendly, comprehensive functional annotation pipeline for microbial genomes". In: *BMC bioinformatics* 22.1 (2021), pp. 1–16.

[280] Daniel Ryan et al. “A high-resolution transcriptome map identifies small RNA regulation of metabolism in the gut microbe Bacteroides thetaiotaomicron”. In: *Nature communications* 11.1 (2020), p. 3557.

[281] Milton H Saier Jr et al. “The transporter classification database (TCDB): 2021 update”. In: *Nucleic acids research* 49.D1 (2021), pp. D461–D467.

[282] R Sangeetha, Arulpandi Geetha, and I Arulpandi. “Optimization of protease and lipase production by Bacillus pumilus SG 2 isolated from an industrial effluent”. In: *Internet J Microbiol* 5.2 (2008), pp. 1–8.

[283] Tsutomu Sato and Yasuo Kobayashi. “The ars operon in the skin element of Bacillus subtilis confers resistance to arsenate and arsenite”. In: *Journal of bacteriology* 180.7 (1998), pp. 1655–1661.

[284] Michael Sauer et al. “The efficient clade: lactic acid bacteria for industrial chemical production”. In: *Trends in biotechnology* 35.8 (2017), pp. 756–769.

[285] Marcus Schallmey, Ajay Singh, and Owen P Ward. “Developments in the use of Bacillus species for industrial production”. In: *Canadian journal of microbiology* 50.1 (2004), pp. 1–17.

[286] Dominique Schneider and Richard E Lenski. “Dynamics of insertion sequence elements during experimental evolution of bacteria”. In: *Research in Microbiology* 155.5 (2004), pp. 319–327.

[287] Amanda N Scholes and Jeffrey A Lewis. “Comparison of RNA isolation methods on RNA-Seq: implications for differential expression and meta-analyses”. In: *BMC genomics* 21.1 (2020), pp. 1–9.

[288] Romy Scholz et al. “Plantazolicin, a novel microcin B17/streptolysin S-like natural product from Bacillus amyloliquefaciens FZB42”. In: *Journal of bacteriology* 193.1 (2011), pp. 215–224.

[289] Tilman Schultze et al. “Current status of antisense RNA-mediated gene regulation in Listeria monocytogenes”. In: *Frontiers in Cellular and Infection Microbiology* 4 (2014).

[290] Torsten Seemann. “Prokka: rapid prokaryotic genome annotation”. In: *Bioinformatics* 30.14 (2014), pp. 2068–2069.

[291] Huanhuan Shao et al. “Construction of novel shuttle expression vectors for gene expression in Bacillus subtilis and Bacillus pumilus”. In: *The Journal of General and Applied Microbiology* 61.4 (2015), pp. 124–131.

[292] Jason W Shapiro and Catherine Putonti. “Gene co-occurrence networks reflect bacteriophage ecology and evolution”. In: *MBio* 9.2 (2018), e01870–17.

[293] Archana Sharma and T Satyanarayana. “Comparative genomics of Bacillus species and its relevance in industrial microbiology”. In: *Genomics insights* 6 (2013), GEI–S12732.

[294] Yishai Shimoni et al. “Regulation of gene expression by small non-coding RNAs: a quantitative view”. In: *Molecular systems biology* 3.1 (2007), p. 138.

[295] Yuki Shimoya. *COGclassifier: A tool for classifying prokaryote protein sequences into COG functional category*. Version 1.2.0. Mar. 20, 2022. URL: github.com/moshi4/COGclassifier.

[296] Yuki Shimoya. *MGCplotter: Microbial Genome Circular plotting tool*. Version 1.2.0. Apr. 10, 2022. URL: github.com/moshi4/MGCplotter.

[297] Patricia Siguier, Jonathan Filée, and Michael Chandler. "Insertion sequences in prokaryotic genomes". In: *Current opinion in microbiology* 9.5 (2006), pp. 526–531.

[298] Gregory T Smaldone et al. "A global investigation of the Bacillus subtilis iron-sparing response identifies major changes in metabolism". In: *Journal of bacteriology* 194.10 (2012), pp. 2594–2605.

[299] Nathan Ryno Smith, Ruth Evelyn Gordon, and Francis Eugene Clark. *Aerobic mesophilic sporeforming bacteria*. Vol. 552. US Department of Agriculture, 1946.

[300] Keunhong Son et al. "A simple guideline to assess the characteristics of RNA-Seq data". In: *BioMed Research International* 2018 (2018).

[301] Rotem Sorek and Pascale Cossart. "Prokaryotic transcriptomics: a new view on regulation, physiology and pathogenicity". In: *Nature Reviews Genetics* 11.1 (2010), pp. 9–16.

[302] John Spizizen. "Transformation of biochemically deficient strains of Bacillus subtilis by deoxyribonucleate". In: *Proceedings of the National Academy of Sciences of the United States of America* 44.10 (1958), p. 1072.

[303] Ilona Stefańska et al. "Antimicrobial susceptibility of lactic acid bacteria strains of potential use as feed additives-the basic safety and usefulness criterion". In: *Frontiers in Veterinary Science* 8 (2021).

[304] Lincoln Stein. "Genome annotation: from sequence to biology". In: *Nature reviews genetics* 2.7 (2001), pp. 493–503.

[305] Jörg Stülke and Larissa Krüger. "Cyclic di-AMP signaling in bacteria". In: *Annu Rev Microbiol* 74 (2020), pp. 159–179.

[306] R Subramaniam et al. "High-density cultivation in the production of microbial products". In: *Chemical and biochemical engineering quarterly* 32.4 (2018), pp. 451–464.

[307] Jing Sun et al. "CodY, ComA, DegU and Spo0A controlling lipopeptides biosynthesis in Bacillus amyloliquefaciens fmbJ". In: *Journal of Applied Microbiology* 131.3 (2021), pp. 1289–1304.

[308] Weifeng Sun et al. "An auto-inducible expression and high cell density fermentation of beefy meaty peptide with Bacillus subtilis". In: *Bioprocess and Biosystems Engineering* 43.4 (2020), pp. 701–710.

[309] Shizhe Tang et al. "Overexpression of an endogenous raw starch digesting mesophilic α-amylase gene in Bacillus amyloliquefaciens Z3 by in vitro methylation protocol". In: *Journal of the Science of Food and Agriculture* 100.7 (2020), pp. 3013–3023.

[310] Roman L Tatusov, Eugene V Koonin, and David J Lipman. "A genomic perspective on protein families". In: *Science* 278.5338 (1997), pp. 631–637.

[311] Roman L Tatusov et al. "The COG database: a tool for genome-scale analysis of protein functions and evolution". In: *Nucleic acids research* 28.1 (2000), pp. 33–36.

[312] Tatiana Tatusova et al. "NCBI prokaryotic genome annotation pipeline". In: *Nucleic acids research* 44.14 (2016), pp. 6614–6624.

[313] Ozlem Tepe and Arzu Y Dursun. "Exo-pectinase production by Bacillus pumilus using different agricultural wastes and optimizing of medium components using response surface methodology". In: *Environmental Science and Pollution Research* 21.16 (2014), pp. 9911–9920.

[314] Felix Teufel et al. "SignalP 6.0 predicts all five types of signal peptides using protein language models". In: *Nature biotechnology* (2022), pp. 1–3.

[315] Santosh Thapa et al. "Biochemical characteristics of microbial enzymes and their significance from industrial perspectives". In: *Molecular biotechnology* 61.8 (2019), pp. 579–601.

[316] Kenneth Timmis et al. "The urgent need for microbiology literacy in society". In: *Environmental Microbiology* 21.5 (2019).

[317] Madhan R Tirumalai et al. "Candidate genes that may be responsible for the unusual resistances exhibited by Bacillus pumilus SAFR-032 spores". In: *PLoS One* 8.6 (2013), e66012.

[318] Harold Tjalsma et al. "Signal peptide-dependent protein transport in Bacillus subtilis: a genome-based survey of the secretome". In: *Microbiology and molecular biology reviews* 64.3 (2000), pp. 515–547.

[319] Shirlley Elizabeth Martínez Tolibia et al. "Engineering of global transcription factors in Bacillus, a genetic tool for increasing product yields: a bioprocess overview". In: *World Journal of Microbiology and Biotechnology* 39.1 (2023), pp. 1–21.

[320] Alexandra Tsirigotaki et al. "Protein export through the bacterial Sec pathway". In: *Nature Reviews Microbiology* 15.1 (2017), pp. 21–36.

[321] Ashley T Tucker et al. "A DNA mimic: The structure and mechanism of action for the anti-repressor protein AbbA". In: *Journal of molecular biology* 426.9 (2014), pp. 1911–1924.

[322] Inam Ul Haq, Sabine Brantl, and Peter Müller. "A new role for SR1 from Bacillus subtilis: regulation of sporulation by inhibition of kinA translation". In: *Nucleic Acids Research* 49.18 (2021), pp. 10589–10603.

[323] Inam Ul Haq, Peter Müller, and Sabine Brantl. "Intermolecular communication in Bacillus subtilis: RNA-RNA, RNA-protein and small protein-protein interactions". In: *Frontiers in Molecular Biosciences* 7 (2020), p. 178.

[324] Taylor B Updegrove, Svetlana A Shabalina, and Gisela Storz. "How do base-pairing small RNAs evolve?" In: *FEMS microbiology reviews* 39.3 (2015), pp. 379–391.

[325] Parag A Vaishampayan et al. "Survival of Bacillus pumilus spores for a prolonged period of time in real space conditions". In: *Astrobiology* 12.5 (2012), pp. 487–497.

[326] Abbe N Vallejo, Robert J Pogulis, and Larry R Pease. "In vitro synthesis of novel genes: mutagenesis and recombination by PCR." In: *Genome research* 4.3 (1994), S123–S130.

[327] Arnoud HM Van Vliet. "Next generation sequencing of microbial transcriptomes: challenges and opportunities". In: *FEMS microbiology letters* 302.1 (2010), pp. 1–7.

[328] Joachim Vandecraen et al. "The impact of insertion sequences on bacterial genome plasticity and adaptability". In: *Critical reviews in microbiology* 43.6 (2017), pp. 709–730.

[329] Kommireddy Vasu and Valakunja Nagaraja. "Diverse functions of restriction-modification systems in addition to cellular defense". In: *Microbiology and molecular biology reviews* 77.1 (2013), pp. 53–72.

[330] Birgit Veith et al. “The complete genome sequence of Bacillus licheniformis DSM13, an organism with great industrial potential”. In: *Microbial Physiology* 7.4 (2004), pp. 204–211.

[331] Pauli Virtanen et al. “SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python”. In: *Nature Methods* 17 (2020), pp. 261–272. DOI: 10.1038/s41592-019-0686-2.

[332] Antti Vuolanto et al. “Phytase production by high cell density culture of recombinant Bacillus subtilis”. In: *Biotechnology Letters* 23.10 (2001), pp. 761–766.

[333] E Gerhart H Wagner and Pascale Romby. “Small RNAs in bacteria and archaea: who they are, what they do, and how they do it”. In: *Advances in genetics* 90 (2015), pp. 133–208.

[334] HY Wang et al. “Screening and mutagenesis of a novel Bacillus pumilus strain producing alkaline protease for dehairing”. In: *Letters in applied microbiology* 44.1 (2007), pp. 1–6.

[335] Liguo Wang, Shengqin Wang, and Wei Li. “RSeQC: quality control of RNA-seq experiments”. In: *Bioinformatics* 28.16 (2012), pp. 2184–2185.

[336] Liguo Wang et al. “Measure transcript integrity using RNA-seq data”. In: *BMC bioinformatics* 17.1 (2016), pp. 1–16.

[337] Qian Wang et al. “Comparative genomic analyses reveal genetic characteristics and pathogenic factors of Bacillus pumilus HM-7”. In: *Frontiers in Microbiology* 13 (2022), p. 4275.

[338] Yi Wang et al. “Deleting multiple lytic genes enhances biomass yield and production of recombinant proteins by Bacillus subtilis”. In: *Microbial cell factories* 13.1 (2014), pp. 1–11.

[339] Bianca Waschkau et al. “Generation of readily transformable Bacillus licheniformis mutants”. In: *Applied microbiology and biotechnology* 78.1 (2008), pp. 181–188.

[340] Bridget NJ Watson, Raymond HJ Staals, and Peter C Fineran. “CRISPR-Cas-mediated phage resistance enhances horizontal gene transfer by transduction”. In: *MBio* 9.1 (2018), e02406–17.

[341] Tilmann Weber and Hyun Uk Kim. “The secondary metabolite bioinformatics portal: computational tools to facilitate synthetic biology of secondary metabolite production”. In: *Synthetic and Systems Biotechnology* 1.2 (2016), pp. 69–79.

[342] Maren Wehrs et al. “Engineering robust production microbes for large-scale cultivation”. In: *Trends in microbiology* 27.6 (2019), pp. 524–537.

[343] Liping Wei et al. “Comparative genomics approaches to study organism similarities and differences”. In: *Journal of biomedical informatics* 35.2 (2002), pp. 142–150.

[344] Stephanie Wemhoff and Friedhelm Meinhardt. “Generation of biologically contained, readily transformable, and genetically manageable mutants of the biotechnologically important Bacillus pumilus”. In: *Applied microbiology and biotechnology* 97.17 (2013), pp. 7805–7819.

[345] Marian Wenzel et al. “Self-inducible Bacillus subtilis expression system for reliable and inexpensive protein production by high-cell-density fermentation”. In: *Applied and environmental microbiology* 77.18 (2011), pp. 6419–6425.

[346] Laura Wicke et al. “Introducing differential RNA-seq mapping to track the early infection phase for Pseudomonas phage KZ”. In: *RNA biology* 18.8 (2021), pp. 1099–1110.

[347] Sandra Wiegand et al. “RNA-Seq of Bacillus licheniformis: active regulatory RNA features expressed within a productive fermentation”. In: *BMC genomics* 14.1 (2013), pp. 1–20.

[348] Julia Wong et al. “Systems-Level Analysis of Bacterial Regulatory Small RNA Networks”. In: *Systems Biology* (2018), pp. 97–127.

[349] Sven Wydra. “Value chains for industrial biotechnology in the bioeconomy-innovation system analysis”. In: *Sustainability* 11.8 (2019), p. 2435.

[350] Liming Xia et al. “Biosynthetic gene cluster profiling predicts the positive association between antagonism and phylogeny in Bacillus”. In: *Nature communications* 13.1 (2022), pp. 1–11.

[351] Zhiqun Xie and Haixu Tang. “ISEScan: automated identification of insertion sequence elements in prokaryotic genomes”. In: *Bioinformatics* 33.21 (2017), pp. 3340–3347.

[352] Bingyue Xin et al. “The Bacillus cereus group is an excellent reservoir of novel lanthipeptides”. In: *Applied and environmental microbiology* 81.5 (2015), pp. 1765–1774.

[353] Yunfan Xu et al. “RNA sequencing reveals small RNAs in Bacillus pumilus under different growth phases of the protease fermentation process”. In: *Applied microbiology and biotechnology* 104.2 (2020), pp. 833–852.

[354] Dan Xue et al. “Correlational networking guides the discovery of unclustered lanthipeptide protease-encoding genes”. In: *Nature communications* 13.1 (2022), pp. 1–14.

[355] Montarop Yamabhai et al. “Secretion of recombinant Bacillus hydrolytic enzymes using Escherichia coli expression systems”. In: *Journal of Biotechnology* 133.1 (2008), pp. 50–57.

[356] Junya Yamamoto et al. “Constitutive expression of the global regulator AbrB restores the growth defect of a genome-reduced Bacillus subtilis strain and improves its metabolite production”. In: *DNA Research* 29.3 (2022), dsac015.

[357] Koichi Yano et al. “Multiple rRNA operons are essential for efficient cell growth and sporulation as well as outgrowth in Bacillus subtilis”. In: *Microbiology* 159.Pt_11 (2013), pp. 2225–2236.

[358] Kazumasa Yasui et al. “Improvement of bacterial transformation efficiency using plasmid artificial modification”. In: *Nucleic acids research* 37.1 (2009), e3–e3.

[359] Kah Yen Claire Yeak et al. “Lichenysin Production by Bacillus licheniformis Food Isolates and Toxicity to Human Cells.” In: *Frontiers in microbiology* 13 (2022), pp. 831033–831033.

[360] Ana Yepes et al. “The biofilm formation defect of a B acillus subtilis flotillin-defective mutant involves the protease FtsH”. In: *Molecular microbiology* 86.2 (2012), pp. 457–471.

[361] Yanglei Yi and Oscar P Kuipers. “Development of an efficient electroporation method for rhizobacterial Bacillus mycoides strains”. In: *Journal of microbiological methods* 133 (2017), pp. 82–86.

[362] Naoto Yoshida and Misa Sato. "Plasmid uptake by bacteria: a comparison of methods and efficiencies". In: *Applied microbiology and biotechnology* 83.5 (2009), pp. 791–798.

[363] Sung-Huan Yu, Jörg Vogel, and Konrad U Förstner. "ANNOgesic: a Swiss army knife for the RNA-seq based annotation of bacterial/archaeal genomes". In: *Gigascience* 7.9 (2018), giy096.

[364] Evgeni M Zdobnov and Rolf Apweiler. "InterProScan–an integration platform for the signature-recognition methods in InterPro". In: *Bioinformatics* 17.9 (2001), pp. 847–848.

[365] Stephan Zellmeier et al. "The absence of FtsH metalloprotease activity causes overexpression of the σW-controlled pbpE gene, resulting in filamentous growth of Bacillus subtilis". In: *Journal of bacteriology* 185.3 (2003), pp. 973–982.

[366] Qingchao Zeng et al. "Comparative genomic and functional analyses of four sequenced Bacillus cereus genomes reveal conservation of genes relevant to plant-growth-promoting traits". In: *Scientific reports* 8.1 (2018), pp. 1–10.

[367] Guo-qiang Zhang et al. "Enhancing electro-transformation competency of recalcitrant Bacillus amyloliquefaciens by combining cell-wall weakening and cell-membrane fluidity disturbing". In: *Analytical biochemistry* 409.1 (2011), pp. 130–137.

[368] Guoqiang Zhang et al. "A mimicking-of-DNA-methylation-patterns pipeline for overcoming the restriction barrier of bacteria". In: (2012).

[369] Kang Zhang, Lingqia Su, and Jing Wu. "Enhanced extracellular pullulanase production in Bacillus subtilis using protease-deficient strains and optimal feeding". In: *Applied microbiology and biotechnology* 102.12 (2018), pp. 5089–5103.

[370] Z Zhang et al. "Improvement of iturin A production in Bacillus subtilis ZK 0 by overexpression of the comA and sigA genes". In: *Letters in Applied Microbiology* 64.6 (2017), pp. 452–458.

[371] Zhi Zhang et al. "Development of an efficient electroporation method for iturin A-producing Bacillus subtilis ZK". In: *International Journal of Molecular Sciences* 16.4 (2015), pp. 7334–7351.

[372] Ziqiang Zheng et al. "The CRISPR-Cas systems were selectively inactivated during evolution of Bacillus cereus group for adaptation to diverse environments". In: *The ISME journal* 14.6 (2020), pp. 1479–1493.

[373] Cuixia Zhou et al. "Optimization of alkaline protease production by rational deletion of sporulation related genes in Bacillus licheniformis". In: *Microbial cell factories* 18.1 (2019), pp. 1–12.

[374] Cuixia Zhou et al. "Spo0A can efficiently enhance the expression of the alkaline protease gene aprE in Bacillus licheniformis by specifically binding to its regulatory region". In: *International journal of biological macromolecules* 159 (2020), pp. 444–454.