PHYLOGENY-ONTOGENY CORRESPONDENCE IN Bacillus subtilis BIOFILMS

Koska, Sara

University of Zagreb

Faculty of Science

Department of Biology

Sara Koska

PHYLOGENY-ONTOGENY CORRESPONDENCE IN *Bacillus subtilis* BIOFILMS

DOCTORAL THESIS

Supervisor:

Dr. rer. nat. Tomislav Domazet-Lošo, Associate Professor

Zagreb, 2022.

Sveučilište u Zagrebu

Prirodoslovno-matematički fakultet

Biološki odsjek

Sara Koska

SUODNOS FILOGENIJE I ONTOGENIJE BIOFILMOVA BAKTERIJE *Bacillus subtilis*

DOKTORSKI RAD

Mentor:

izv. prof. dr. rer. nat. Tomislav Domazet-Lošo

Zagreb, 2022.

This doctoral thesis has been produced by the author in the Laboratory of Evolutionary Genetics, Division of Molecular Biology, Ruđer Bošković Institute, under the supervision of dr. rer. nat. Tomislav Domazet-Lošo, Associate Professor, as a part of the University postgraduate study of Biology at the Division of Biology, Faculty of Science, University of Zagreb.

Ovaj je doktorski rad izrađen u Laboratoriju za evolucijsku genetiku Zavoda za molekularnu biologiju Instituta Ruđer Bošković, pod vodstvom dr. rer. nat. Tomislava Domazeta-Loše, izv. prof., u sklopu Sveučilišnog posijediplomskog doktorskog studija Biologije pri Biološkom odsjeku Prirodoslovno-matematičkog fakulteta Sveučilišta u Zagrebu.

SUPERVISOR INFORMATION

Tomislav Domazet-Lošo graduated biology at the University of Zagreb (1997) and earned a doctorate in evolutionary genetics from University of Cologne (2003). From 2008 to 2011 he held postdoctoral positions at the Max-Plank Institute for Evolutionary biology and Zoological Institute, University of Kiel, Germany. Currently, he is a group leader at the Ruđer Bošković Institute and associate professor at the School of Medicine, Catholic University of Croatia in Zagreb.

His research focuses on the macroevolutionary dynamics across the tree of life. He combines experiments and computational approaches to get deeper understanding of the origin of new genes, evolution of development and evolution of biological complexity. In his latest research, he used phylostratigraphic technology to address evolutionary and functional questions in bacteria, especially those related to the evolution of bacterial multicellular behavior.

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my supervisor dr. rer. nat. Tomislav Domazet-Lošo for giving me the opportunity to be a part of his group in the Laboratory for evolutionary genetics, and for generously sharing his knowledge and support throughout my training as a PhD student.

I thank all who participated in the story about ontogeny-phylogeny correlation in *Bacillus subtilis* biofilms. The making and completion of my PhD would not have been possible without each of their important contributions.

My gratitude also goes to all the teachers and professors who believed in me and encouraged me to push harder and achieve more than I believed I could throughout my education.

Finally, many thanks to my family and my friends for their understanding and support in everything.

Sveučilište u Zagrebu
Prirodoslovno-matematički fakultet
Biološki odsjek

Doktorski rad

SUODNOS FILOGENIJE I ONTOGENIJE BIOFILMOVA BAKTERIJE *Bacillus subtilis*

SARA KOSKA

Institut Ruđer Bošković

Korelacija između filogenije i ontogenije životinja je predmet rasprave među znanstvenicima od početka 19. stoljeća. Razvojem molekularne biologije i bioinformatike te analizama transkriptoma razvojnih stadija, uočena je prisutnost korelacije filogenije i ontogenije na molekularnoj razini u eukariotskim evolucijskim granama sa složenom višestaničnosti. Ova saznanja potaknula su nas da testiramo korelaciju filogenije i ontogenije u bazalnim organizmima s nejasnim razvojnim i višestaničnim karakteristikama. Analizom transkriptomskih i proteomskih podataka, pokazali smo da ontogenija *Bacillus subtilis* biofilmova korelira s evolucijskim mjerama u obliku rekapitulacijskog profila, odnosno evolucijski mlađi i divergentniji geni su sve više eksprimirani prema kasnijim stadijima rasta biofilma. Dodatne molekularne analize pokazale su da je rast biofilma reguliran i organiziran u jasno odvojene stadije ontogenije. Zaključno, rezultati pokazuju da je rast *Bacillus subtilis* biofilmova istinski razvojni proces usporediv s razvojem životinja, biljaka i gljiva.

Ključne riječi: *biofilm, Bacillus subtilis, korelacija filogenije i ontogenije, transkriptom, proteom, evolucijska razvojna biologija*
Mentor: dr. rer. nat. Tomislav Domazet-Lošo, izv. prof.
Ocjenjivači: izv. prof. dr. sc. Damjan Franjević, PMF, Zagreb
prof. dr. sc. Jasna Hrenović, PMF, Zagreb
prof. dr. sc. Ivan Mijaković, Chalmers University of Technology, Švedska
Zamjena: prof. dr. sc. Đurđica Ugarković, IRB, Zagreb

University of Zagreb
Faculty of Science
Department of Biology

Doctoral thesis

PHYLOGENY-ONTOGENY CORRESPONDENCE IN *Bacillus subtilis* BIOFILMS

SARA KOSKA

Ruđer Bošković Institute

Links between ontogeny and phylogeny in animals have been discussed for more than two centuries. With the uprising of molecular biology and bioinformatics, several studies have revealed the presence of the phylogeny-ontogeny correlation on molecular level in developmental transcriptomes of eukaryotic clades with complex multicellularity. These findings open a possibility to test the phylogeny-ontogeny correlation in more basal organisms, with more obscure development and multicellularity characteristics. Using time-resolved transcriptome and proteome profiles, this study showed that *Bacillus subtilis* biofilm ontogeny correlates with the evolutionary measures through recapitulation pattern, in a way that evolutionary younger and more diverged genes were increasingly expressed towards the later timepoints of the biofilm growth. Molecular and morphological signatures also revealed that biofilm growth is highly regulated and organized into discrete ontogenetic stages. Together, this suggests that the biofilm formation in *Bacillus subtilis* is a true developmental process comparable to organismal development in animals, plants and fungi.

Keywords: *biofilms, Bacillus subtilis, phylogeny-ontogeny correlations, transcriptome, proteome, evo-devo*
Supervisor: dr. rer. nat. Tomislav Domazet-Lošo, Associate Professor
Reviewers: dr. sc. Damjan Franjević, Associate Professor, PMF, Zagreb
dr. sc. Jasna Hrenović, Professor, PMF, Zagreb
dr. sc. Ivan Mijaković, Professor, Chalmers University of Technology, Sweden
Replacement: dr. sc. Đurđica Ugarković, Professor, IRB, Zagreb

TABLE OF CONTENTS

1. INTRODUCTION

The discussion about phylogeny-ontogeny relation in the animal kingdom has been going on for almost 200 years. Many of the proposed models about this correspondence emphasize the idea of their parallelism, based on similarities between embryos of different species in the same animal phylum, their morphological complexity, and the fossil record (Abzhanov, 2013). One of the best known models is the one by Ernst Haeckel (Haeckel, 1868), shortly accepted as "ontogeny recapitulates phylogeny". Opposite to recapitulation models, Karl Ernst von Baer (Baer, 1828) proposed his "law of embryology" which connects the increase in complexity of adult animal forms with the formation of their traits during embryogenesis (Abzhanov, 2013). Some of the derivations from von Baer's law of embryology include the so-called "funnel" model and the "hourglass" model with the concept of phylotypic stage, which shapes the narrowest part of the hourglass model and presents a striking morphogenetic resemblance in different species of a certain phylum (Slack, Holland and Graham, 1993; Duboule, 1994; Kalinka and Tomancak, 2012; Abzhanov, 2013).

With the uprising of methods in molecular biology and bioinformatics, several studies have confirmed the developmental hourglass model on molecular level, which underpins morphological hourglass pattern not just in animals, but in various multicellular taxa (Domazet-Lošo and Tautz, 2010; Kalinka *et al.*, 2010; Quint *et al.*, 2012; Cheng *et al.*, 2015). For example, Kalinka *et al.* (2010) have shown in their study that gene expression is maximally conserved during the arthropod phylotypic period, meaning that gene expression is more resistant to evolutionary changes during mid-development compared to either early or late development. Domazet-Lošo and Tautz (2010) have also confirmed the molecular hourglass on arthropods and other animal phyla, based on their method phylostratigraphy that combines gene expression level with gene evolutionary history (Domazet-Lošo, Brajković and Tautz, 2007). Phylostratigraphy was also successful in determining phylogeny-ontogeny correlation in multicellular plants and fungi (Quint *et al.*, 2012; Cheng *et al.*, 2015). Although animals, multicellular plants and fungi developed multicellularity independently (Niklas, 2014), they all show the same pattern of phylogeny-ontogeny correlation, where evolutionary younger and more diverged genes are used during early and late development, while evolutionary older and less diverged genes are used during mid-development (Domazet-Lošo and Tautz, 2010; Quint *et al.*, 2012; Cheng *et al.*, 2015).

Bioinformatic tools that link evolutionary and developmental data can in principle be used on different taxa, but they have not been applied for investigating phylogeny-ontogeny correlation in more basal organisms, such as bacterial biofilms. Bacterial biofilms are communities of genetically identical bacterial cells that express different genetic programs and produce subpopulations of functionally distinct, but coexisting cell types (Vlamakis *et al.*, 2013). Bacteria are widely recognized as simple, unicellular organisms, but several decades ago, some of the researchers started to consider them as multicellular organisms. This opinion in based on observations that bacterial cells communicate among themselves which enables them decision-making and division of labour, coordination of growth, movement and biochemical activities, and many benefits including more efficient proliferation and access to resources otherwise inaccessible, or collective defence and population survival (Shapiro, 1998). Although bacterial biofilms in nature mostly include multiple bacterial species (Røder, Sørensen and Burmølle, 2016), studies on single-species biofilms have been of the most importance in learning and understanding the biofilm biology (Vlamakis *et al.*, 2013). In the period of the last two decades, non-pathogenic and Gram-positive bacterium *Bacillus subtilis* has become known as a model organism for studying biofilm formation.

The aim of this doctoral thesis is to determine the correspondence between phylogeny and ontogeny in bacterial biofilms, traditionally considered as simple organisms, but showing plethora of collective properties. If the expression levels of genes are independent of their evolutionary age and divergence rates, one should expect to find age and divergence indices that show a trend close to a flat line across biofilm developmental timepoints; *i.e.*, there will be no correlation between ontogeny and phylogeny. Conversely, the existence of any correlation between age or divergence indices and biofilm developmental timepoints would indicate that biofilm development harbours macroevolutionary logic similar to embryos of multicellular eukaryotes. The presence of this correlation is until now known for eukaryotic multicellular organisms only. In addition, only transcriptome expression data have been used in studying phylogeny-ontogeny correlations so far, which enables including proteome quantification data in this study. The study was made by bioinformatic analyses of transcriptome expression and proteome quantification data gained by RNA and protein isolation from *Bacillus subtilis* subsp. *subtilis* str. NCIB3610 from ten timepoints covering biofilm growth from its inoculation until two months of culturing.

2. LITERATURE REVIEW

2.1. Phylogeny-ontogeny correlation

2.1.1. Morphological phylogeny-ontogeny correlation in animals

Understanding the relationship between phylogeny and ontogeny is essential in the fields of developmental and evolutionary biology and genetics. First proposed models describing phylogeny-ontogeny correlation in animals date back to 1820s. One such model is by Johann Meckel and Etienne Serres, who were stating that all animals share a universal body form that advances from simple to complex, and that embryonic structures of "higher" animals are comparable with organs in "lower" animals (Abzhanov, 2013). Following theories, like the one by Louis Agassiz, recognized different types of body plans, and expanded the parallelism of ontogeny and phylogeny in "lower" and "higher" animals to fossil records (Abzhanov, 2013). One of the most extreme models describing this parallelism was the "recapitulation theory" or "biogenetic law" by Ernst Haeckel (Haeckel, 1868), shortly explained with the phrase "ontogeny recapitulates phylogeny". Haeckel took into consideration novel evolutionary concepts by Charles Darwin and Alfred Russell Wallace, and stated that the organism during its development recapitulates the evolution of its species, *i.e.* the size and the shape changes during organismal ontogeny recapitulate its species phylogeny (Abzhanov, 2013). Haeckel claimed that the development of higher animal goes through the stages that correspond to the adult organism of the more primitive animal (Abzhanov, 2013). One of the first proposed models was also the one by Karl Ernst von Baer (1828), but it was opposite to then widely accepted recapitulation models. Von Baer proposed his what is today known as "developmental law" or "law of embryology". He observed some general trends during animal embryonic development – general attributes of a larger group of animals appear earlier than the special characters in their embryos; special forms are developed from the general forms; every embryo of a given animal form becomes separate from the other forms and does not pass through them; and the embryo of a higher form resembles only its embryo, not any other form (Baer, 1828). Law of embryology states that tissue and organ differentiation increase during development and specialized structures arise from the general, so this rules out recapitulation, as the embryo of a higher animal does not represent the adult of a lower animal (Abzhanov, 2013). Today, Haeckel's idea that animal embryology is a recurrence of their evolution is widely rejected, and von Baer's laws are mostly accepted. One of the most famous derivations of von Baer's third

law are the so-called “funnel” model and the “hourglass” model with the concept of phylotypic period. The “funnel” model, also known as the “early conservation” model, implies that the embryos are highly conserved in their early stages, and progressively diverge and branch as development advances (Abzhanov, 2013). It assumes that mutations or perturbations at the earlier stages of development can have a widespread effects at the following developmental stages, thus they have to be conserved (Irie and Kuratani, 2011). On the contrary, the phylotypic period of the “hourglass” model presents a part of mid-embryonic development when the basic body plan of a certain phylum is being established with all major body parts represented in their final positions and which is morphogenetically conserved due to natural selection or developmental constraints in different species (Slack, Holland and Graham, 1993; Duboule, 1994; Kalinka and Tomancak, 2012; Abzhanov, 2013). One hypothetical motive for mid-embryonic conservation is the activation of spatial and temporal collinearity of *Hox* genes during that stage, which make the animal zootype and are essential for the proper body plan organization during development (Slack, Holland and Graham, 1993; Duboule, 1994). However, even animals without the *Hox* cluster have a mid-development transition period which corresponds to a phylotypic period (Levin *et al.*, 2016). The other hypothesis suggests that networks of local and global inductive signals are responsible for conservation and developmental interdependence of different organ primordia and therefore any changes during mid-development stages would increase the risk of mortality (Irie and Kuratani, 2011). It is important to note that the term “early embryogenesis” today refers to the stages during which blastomere cleavage, blastulation, gastrulation, and early neurulation occur, and they are considered to be a part of the adult maternal phenotype which evolved more recently and can differ among closely related species due to environmental conditions (Abzhanov, 2013). In von Baer’s time, the pharyngula stage was considered as a starting point of the “early embryogenesis”, and he observed the embryos from that point onward (Abzhanov, 2013). Regardless of different opinions relating to starting, mid or terminal stages of embryogenesis, it is clear that at some timeframe during animal embryogenesis there are striking morphological similarities in the embryos across different species in a certain phylum (Slack, Holland and Graham, 1993; Duboule, 1994; Abzhanov, 2013)

2.1.2. Molecular phylogeny-ontogeny correlation

With the uprising of methods in molecular biology and bioinformatics, more and more studies that tried to explore the relationship between evolutionary and developmental processes on the molecular level started to emerge. Although there is a general accordance with the developmental hourglass considering animal morphology, there are some disagreements in molecular studies. For example, Roux and Robinson-Rechavi (2008) found that essential genes and genes with strong sequence constraints in zebrafish and mouse tend to be expressed in early developmental stages in comparison to late developmental stages, thus confirming the early conservation model on molecular level. Comte, Roux and Robinson-Rechavi (2010) also confirmed the molecular early conservation model, by showing that early expressed genes are more conserved between zebrafish and mice and are regulated by different pathways compared to genes expressed during later developmental stages. The authors hypothesized that reduced phenotypic diversity during the phylotypic period can be explained with the conservation in mechanism of pattern formation at the earlier stages (Richardson, 1999). Also, divergence following the phylotypic period can be assigned to changes in developmental mechanisms during that period, and these changes are expressed and enlarged later in development (Richardson, 1999). Levin *et al.* (2016) used the approach of comparative transcriptomics and found that orthologous gene expression is comparable between sets of ten different species from ten different phyla in early and late development, indicating a molecular reverse hourglass model. Between the periods of cell proliferation in early development and differentiation in late development that are similar between phyla, a period of transition during mid-development occurs where signalling pathways and transcription factors that vary between phyla are enriched (Levin *et al.*, 2016). Although the between phyla transcription variance points to the reverse hourglass, the mid-development transitions correspond to the phylotypic period of an hourglass shape within an individual phylum (Levin *et al.*, 2016). Kalinka *et al.* (2010) were also using comparative transcriptomics and measuring gene expression conservation during development of six *Drosophila* species, and found that gene expression is more resistant to evolutionary change during mid-embryogenesis compared to either early or late periods of embryogenesis, meaning that selective constraint is maximized by natural selection during the arthropod phylotypic period. The authors also found that genes contributing the most to the hourglass model are involved in core developmental processes (Kalinka *et al.*, 2010). Similarly, Irie and Kuratani (2011) discovered that gene expression profiles of four vertebrate species have the highest conservation in pharyngula embryos. They found that the genes with conserved

expression include *Hox* genes, genes involved in cell-cell signalling, transcription factors and morphogens or growth factors (Irie and Kuratani, 2011). Also, the genes with similar expression profiles have the higher proportion of development-related genes than the ones with different expressions (Irie and Kuratani, 2011). In a study that had a different approach by using genomic phylostratigraphy (Domazet-Lošo, Brajković and Tautz, 2007), Domazet-Lošo & Tautz (2010) linked evolutionary gene age with gene expression levels and calculated transcriptome age index (TAI) for 60 stages covering zebrafish ontogeny. Their results showed that the TAI has the lowest value during the late segmentation/early pharyngula stage, meaning that evolutionary oldest transcripts are expressed at that moment, which corresponds to the phylotypic stage in zebrafish. They suggested that adaptations occur primarily as a results of variable environments and juveniles and adults interact much more with ecological factors in comparison with embryos. Mid-embryonic stages around the phylotypic period are not in direct contact with the environmental conditions, and thus have low morphological and molecular divergence (Domazet-Lošo and Tautz, 2010). The authors also confirmed the phenomenon of preferential expression of evolutionary older genes during mid-development in *Drosophila melanogaster*, *Caenorhabditis elegans*, and *Anopheles gambiae* (Domazet-Lošo and Tautz, 2010). Recognizing the potential of genomic phylostratigraphy, Quint *et al.* (2012) wanted to see if molecular hourglass is detectable in another living kingdom that has evolved embryogenesis. Plants do not show discernible morphological hourglass, yet the hourglass pattern has been confirmed on molecular level in *Arabidopsis* (Quint *et al.*, 2012). Moreover, plants exhibit differences in developmental processes during embryogenesis in comparison with metazoa, as animal development mostly occurs during embryogenesis and simultaneously, while majority of plant organs develop post-embryonically and sequentially (Drost *et al.*, 2017). Except confirming that evolutionary the oldest genes during plant gametic embryogenesis are expressed during mid-development, the authors also discovered that the least diverged genes are also preferentially expressed in mid-development, at the transition from morphogenesis to growth, in comparison with early and late embryogenesis (Quint *et al.*, 2012). Molecular hourglass regarding gene age and divergence was also confirmed in mushroom-forming fungi *Coprinopsis cinerea* development, what is even more surprising, since there is no embryogenesis in mushroom like in animals and plants (Cheng *et al.*, 2015). Additionally, although in flowering plants and fungi the "waist" in molecular hourglass model occurs after what is considered to be analogous to phylotypic period in animals, all three mentioned major taxa still exhibit the same molecular phylogeny-ontogeny correlation (Quint *et al.*, 2012; Cheng *et al.*, 2015). Besides corroborating the molecular hourglass during plant embryogenesis, Drost

et al. (2016) found the same pattern even outside the period during which the organogenesis occurs, namely in the transition from embryonic to vegetative phase, and in the transition from a vegetative to a reproductive phase. Drost *et al.* (2017) argue that all developmental processes pass through an organizational checkpoint between two major sequential developmental programs by preventing a larger period of their overlapping. This way, a successful and ordered transition to the successive developmental phase is ensured. The nonexistence of morphological hourglass in plants and fungi also questions the putatively causal relationship between morphological and molecular hourglass patterns, and it is likely irrelevant whether the convincing molecular pattern penetrates or not to the morphological level (Drost *et al.*, 2017). Although different methods draw different conclusions about phylogeny-ontogeny correlations on molecular level in animals, genomic phylostratigraphy reflected the molecular hourglass pattern not only across several animal phyla, but also in two other living kingdoms encompassing eukaryotic multicellular organisms. Despite the fact that animals, plants and fungi developed multicellularity independently (Niklas, 2014), genomic phylostratigraphy was successful in reproducing the same relationship between the organismal development and its evolution in all three above mentioned lineages.

2.2. Bacterial biofilms

Bacterial biofilms are complex biological systems that represent communities of bacterial cells attached to a surface and enclosed in a self-produced extracellular matrix. Aside from bacteria, archaea are another group of organisms capable of forming complex multicellular structures encased in an extracellular matrix. Extracellular matrix is important for keeping the integrity of a biofilm and holding the community together and protecting it (Vlamakis *et al.*, 2008, 2013). It is a common feature of all biofilms, and it is usually composed of a polysaccharide biopolymer along with other components such as proteins or DNA (López, Vlamakis and Kolter, 2010; Vlamakis *et al.*, 2013). Along with bacteria or archaea, biofilms in nature consist additionally of fungi, protozoa, and algae (Romaní *et al.*, 2008). The archaeal and bacterial ability to form biofilms is an ancient and a wide-spread characteristic, regarding that prokaryotic "living fossil" lineages and a wide range of bacterial species are capable of biofilm arrangement and they spend a large part of their lifetime within a biofilm community (Hall-Stoodley, Costerton and Stoodley, 2004; Monds and O'Toole, 2009).

Bacterial biofilms are found on almost all biological and non-biological surfaces and are structurally adapted to survive in varying environments, resistant to different and extreme temperatures, pH, exposures to ultraviolet light, or nutrient conditions (Hall-Stoodley, Costerton and Stoodley, 2004). They occur in almost every environment with sufficient moisture and nutrient amount and on surfaces where their attachment is possible (Singh, Paul and Jain, 2006). Bacterial biofilms are in general considered problematic in many man-made setups, mostly because of their resistance to antibiotic and antimicrobial agents that causes damage in medical and industrial settings (Hall-Stoodley, Costerton and Stoodley, 2004). Furthermore, bacterial biofilm infections are known to affect human teeth, skin, and the urinary tract (López, Vlamakis and Kolter, 2010). This resistance can be explained by the barrier properties of the extracellular matrix, dormant zones with starved and stationary cells within the biofilm, and the existence of resistant phenotypes (Hall-Stoodley, Costerton and Stoodley, 2004). Apart from their negative impact, they can also be exploited beneficially, like a potential source of energy in the form of microbial fuel cells, which can use almost any source of biodegradable organic matter for power generation (Logan, 2009). Their potential in effluent treatment is enormous, as they are more efficient, economic, and safer to use in comparison to chemical or physical methods (Singh, Paul and Jain, 2006). Bacterial biofilms have also been recognized as biological control agents in agriculture, since bacteria in the rhizosphere can form biofilms that help plants in coping with infections from pathogenic bacteria, fungi or nematodes (Vlamakis *et al.*, 2013).

Bacterial biofilms can be grown in different laboratory conditions and thus be in different formations, including colony biofilms at the air-agar interface, floating biofilms at the air-liquid interface, and submerged, surface-adhered biofilms at the liquid-solid interface (Vlamakis *et al.*, 2013). Bacterial biofilms found in nature usually include multiple bacterial species (Røder, Sørensen and Burmølle, 2016), but clinical relevance of biofilms made researchers use model systems that included Gram-negative pathogenic bacteria, like *Pseudomonas aeruginosa* which became the most studied bacterium in the biofilm field (Vlamakis *et al.*, 2013). However, *Bacillus subtilis*, a non-pathogenic and Gram-positive bacterium, specifically an undomesticated wild type strain NCIB3610, has become known as a model organism for studying biofilm formation over the past two decades.

2.2.1. *Bacillus subtilis* biofilm morphology

Bacillus subtilis is a widespread bacterium broadly adapted to many environments. Typical wrinkled morphology of *B. subtilis* biofilms is a consequence of localized cell death and resilience against environmental extremes provided by the extracellular matrix (Asally *et al.*, 2012). The detailed morphological architecture of *B. subtilis* biofilms depends on the used strain and environmental conditions, such as medium composition, incubation temperature, and agar content (Aguilar *et al.*, 2007). At the beginning of biofilm forming, cells are in a shape of short motile rods, and afterwards they become non-motile and form long chains by adhering to each other and to the surface and secrete an extracellular matrix (Vlamakis *et al.*, 2013). Matrix-producing cells become sporulating cells in mature biofilms, while in aged biofilms the extracellular matrix starts to disassemble in response to self-produced D-amino acids and spores can then disperse in the environment (Kolodkin-Gal *et al.*, 2010; Vlamakis *et al.*, 2013).

Vlamakis *et al.* (2008) were determining cell types by monitoring gene expression in colony biofilms and found that gene expression during biofilm formation is temporally and spatially regulated and produces at least three different cell types – motile, matrix-producing and sporulating cells. They found that motile cells, which express *hag* that codes for flagellin, are the most numerous in the early biofilm, at 12 h respectively. They start to decrease in number as the biofilm matures, when matrix-specific genes in matrix-producing cells, like *tapA* encoding for major protein component, peak in the expression at 24 h. Sporulation and *sspB* expression, encoding for protein found only in sporulating cells, starts at 48 h of biofilm formation (Vlamakis *et al.*, 2008).

Vlamakis *et al.* (2008) also determined in what order cell types differentiate. They found that motile cells move to the base and edge of the colony as the biofilm matures. Matrix-producing cells can be found throughout the biofilm in random patches, although they show larger density toward the edge of the colony with biofilm aging. Sporulating cells are preferentially located at the aerial biofilm structures or fruiting bodies. Matrix-producing cells arise from motile cells by repressing their motility, and sporulating cells are mostly derived from the matrix-producing cells. Also, the minority of sporulating cells can arise from motile cells as well (Vlamakis *et al.*, 2008). The process of matrix-producing cells is reversible and can alter as environmental conditions change, so these cells can become motile again (Vlamakis *et al.*, 2008, 2013).

Results from the study by Vlamakis *et al.* (2008) were additionally corroborated in the study by Srinivasan *et al.* (2018). These authors found that the matrix production and the onset of sporulation in growing *B. subtilis* biofilms is localized to the cells within a radially proliferating

front at the biofilm periphery. These fronts represent gene expression waves which move through non-motile bacteria, and do not include cell migration or colony spreading as previously though. The second front arises when cells gradually start to turn off matrix production and turn on sporulation (Srinivasan *et al.*, 2018). Together, these results indicate that biofilm formation is dynamic and complex mechanism which includes different coexisting cell types that change their proportion and localization throughout this process.

2.2.2. Molecular basis for biofilm development in *Bacillus subtilis*

Biofilm formation is a complex process and a result of many interconnected regulatory networks. Although cells within a single-species biofilm are genetically identical as they arise from a single cell, they express different genetic programs that produce subpopulations of functionally and phenotypically distinct, but coexisting cell types (López, Vlamakis and Kolter, 2010; Vlamakis *et al.*, 2013). The process of biofilm formation in *B. subtilis* starts with the expression of genes responsible for matrix production, triggered by an external signal, such as antimicrobial lipoprotein surfactin (Vlamakis *et al.*, 2013).

Central transcription regulator in *B. subtilis* which, among other genes, controls those necessary for biofilm formation and sporulation, is Spo0A. Its activity depends on its direct or indirect phosphorylation by several different kinases, including KinA, KinB, KinC and KinD, in a way that intermediate levels of Spo0A activate matrix gene expression, and higher levels of Spo0A induce sporulation (Vlamakis *et al.*, 2013). Phosphorylated Spo0A controls the activity of the master regulator SinR and AbrB which act repressively on the *eps* and *tapA-sipW-tasA* operons and regulatory gene *slrR* (Kearns *et al.*, 2004; Chu *et al.*, 2008). SlrR and his paralogue SlrA control the biofilm formation by binding to SinR and preventing it from repressing the *eps* and *tapA-sipW-tasA* operons, or by repressing the promoters for motility gene *hag* and genes involved in cell separation *lytABC* and *lytF* (Chai, Kolter and Losick, 2009; Chai *et al.*, 2010; Vlamakis *et al.*, 2013). Phosphorylated Spo0A also controls the SinR antirepressor SinI (Kearns *et al.*, 2004).

Another important transcription regulator in *B. subtilis* is *comA*. It produces protein whose phosphorylation leads to surfactin production via *srf*, which is important in extracellular matrix production (López and Kolter, 2010). Extracellular matrix production is mostly mediated by self-produced extracellular signals, and surfactin is one of the most well-known. Surfactin is a lipopeptide that causes potassium leakage in the membrane, which activates sensor kinase KinC

that triggers phosphorylation of Spo0A and positively regulates matrix gene expression (Lopez *et al.*, 2009; Vlamakis *et al.*, 2013). ComA also activates *degQ* transcription, and DegQ modulates DegU phosphorylation and synthesis of exoproteases and other extracellular enzymes that help with nutrient acquisition (Kobayashi, 2007; Marlow *et al.*, 2014; Spacapan, Danevčič and Mandic-Mulec, 2018). ComX is an autoinducer and quorum-sensing signal responsible for phosphorylation and activation of ComA (Kalamara *et al.*, 2018). Matrix-producing cells secrete the extracellular polysaccharide (EPS) and the structural matrix-associated proteins, which are the main components of the extracellular matrix (López and Kolter, 2010). The major EPS component of all *B. subtilis* biofilms is synthesized by the products of the *epsABCDEFGHIJKLMNO* operon, and two other genes important for the production of nucleotide sugars involved in the *eps* pathway include *pgcA* and *gtaB* (Vlamakis *et al.*, 2013). The two known and characterized proteins present in the matrix are translocation-dependent antimicrobial spore component TasA and biofilm surface layer protein BslA. Functioning of TasA is closely connected to protein TapA, and these two proteins are processed by the third protein type I signal peptidase W SipW, while all three proteins are encoded by the *tapA-sipW-tasA* operon (Vlamakis *et al.*, 2013). Another enzyme important for *B. subtilis* matrix production is KinD, which acts both as phosphatase by keeping low levels of phosphorylated master regulator Spo0A until matrix is sensed, and then starts to phosphorylate Spo0A to promote sporulation (Aguilar *et al.*, 2010).

As biofilm matures, some cells in it eventually become sporulating, which can be triggered by harsh environmental conditions and as a response to natural products from other microorganism, or can be controlled by quorum sensing dependent manner (López and Kolter, 2010; Vlamakis *et al.*, 2013). KinA is a kinase that is responsible for responding to starvation and inducing sporulation through phosphorylation of Spo0A (López and Kolter, 2010). Another mechanism triggering sporulation during starvation is through depletion of GTP which inactivates CodY, which usually serves as an inhibitor of Spo0A expression (Ratnayake-Lecamwasam *et al.*, 2001). Sporulation regulated by quorum sensing includes extracellular peptides Phr that form complex with Rap proteins which stabilize the phosphorylated form of Spo0A and favor sporulation (López and Kolter, 2010). Besides sporulation, nutrient-limited conditions can also lead to the production of an extracellular matrix in *B. subtilis* in a non-signalling mechanism (López and Kolter, 2010). Low levels of phosphorylated Spo0A lead to the expression of two operons encoding for toxin peptides sporulation killing factor SkfA and sporulation-delaying protein SdpC in cannibal cells, which are at the same time resistant to these toxins. As low levels of phosphorylated Spo0A trigger both matrix production and

cannibalism, usually the subpopulations of these two types of cells overlap (López and Kolter, 2010; Vlamakis *et al.*, 2013).

To conclude, biofilm formation is triggered by signals that lead to phosphorylation in regulators that act on a fraction of cells in a culture and are responsible for activation and regulation of processes that result in distinct cell types in *Bacillus subtilis.* This bimodality is achieved when the regulators activate a certain subset of genes until a specific regulator expression level is achieved (López and Kolter, 2010). The size of the subpopulation of cells affected by the regulator also depends on the environmental conditions and signalling molecules, and besides responding to self-produced signals that trigger cell differentiation and biofilm formation, *B. subtilis* can also sense products produced by other bacteria that share its ecological niche (López and Kolter, 2010).

2.2.3. Bacterial biofilms as multicellular organisms and biofilm growth as a developmental process

Bacteria are traditionally considered unicellular organisms as microbiologist used them as single-celled, pure-cultured and domesticated strains in laboratory conditions, which made bacteria lose many of their multicellular attributes (Aguilar *et al.*, 2007). However, there are many behaviours that characterize bacterial biofilms as a multicellular and social organism. One of these behaviours that had a major influence in changing the perspective and made microbiologists start to look at bacteria as multicellular organisms includes communication between cells and diverse signalling networks, which are known to be widespread in biofilms. Production of exoproteases, surfactin and extracellular matrix in fact present a type of intercellular communication in biofilms. These can all be considered as public goods, as they are energetically expensive to produce, but are beneficial to all cells in a bacterial community (Kalamara *et al.*, 2018). Only a subpopulation of cells produces surfactin and expresses matrix genes, and the cells that respond to surfactin are not the ones producing it. This mean of signalling between bacterial cells in which the signalling molecules are self-produced is known as quorum sensing, and if the signal is unidirectional and the producers do not respond to it is referred to as a paracrine signalling (López and Kolter, 2010; Vlamakis *et al.*, 2013). Through this communication, cells within a biofilm can in a way make a decision and adjust its activity in accordance with other cells in a community, which is similar to decision-making capabilities of cells in higher organism (Shapiro, 1998). Communication in *B. subtilis* biofilms can also be

through electrical signalling. The cells from the periphery of biofilm colony periodically halt their growth in order to prevent starvation and death of the cells from the interior, thus providing the overall viability of the biofilm (Liu *et al.*, 2015). This coordination in metabolism is mediated by potassium ion-channel electrical signals (Prindle *et al.*, 2015). To conclude, as Shapiro (1998) states, bacterial multicellularity can be viewed as the ability of individual cell to receive, process and respond to signals from other cells, and this transfer of information between components of the whole system is what makes the notion of an organism.

Spatiotemporal gene expression regulation is also one of the important conditions to meet for the development of multicellular organism, and *B. subtilis* meets that criterion through activating and repressing different genes in several different cell types throughout biofilm formation (Vlamakis *et al.*, 2008; Srinivasan *et al.*, 2018). Changes in spatiotemporal organization are not only detectable on molecular level, but also in changes in morphology and through macroscopic organization that occur during biofilm formation (Monds and O'Toole, 2009). Closely connected to the expression of genes in space and time are gene regulatory networks. They represent another major requirement of a developmental system (Monds and O'Toole, 2009), and there are several known complex regulatory pathways and molecular underpinnings that control biofilm formation in *B. subtilis* (López and Kolter, 2010; Vlamakis *et al.*, 2013), which are described in the introductory part of this thesis.

Development of animals as multicellular organisms appears as punctuated and modular process (Yanai, 2018), and biofilm growth shows some similarities as it can be divided into several disparate stages. Changes in morphology and on molecular level that qualify biofilm growth as a developmental process include the attachment of the cells to a surface, development and maturation of biofilm, and biofilm disassembly and detachment, accompanied with turning-on and turning-off different sets of genes (Stoodley *et al.*, 2002; Vlamakis *et al.*, 2013).

3. MATERIAL AND METHODOLOGY

3.1. Biofilm growth, RNA extraction and protein digestion

Biofilm cultivating and RNA extraction were previously done by the members of the Domazet-Lošo group (Laboratory of Evolutionary Genetics, Department of Molecular Biology, Ruđer Bošković Institute). Briefly, MSgg agar plates were inoculated with *Bacillus subtilis* subsp. *subtilis* str. NCIB3610 (*B. subtilis*) from LB culture. The plates were incubated at 30 °C and the biofilms were harvested for RNA extraction at 6 and 12 hours, and at 1, 2, 3, 5, 7, 14, 30 and 60 days post-inoculation time (transcriptome samples 6H, 12H, 1D, 2D, 3D, 5D, 7D, 14D, 1M and 2M, respectively), and for protein digestion at 12 hours, and at 1, 2, and 7 days post-inoculation time (proteome samples 12H, 1D, 2D and 7D, respectively) (Figure 1a). All samples, excluding 2M (only one replicate due to technically demanding RNA extraction from aged biofilms), were taken in three biological replicates per timepoint. Mass spectrometry analyses of sampled proteins were performed at the Proteomics Core Facility, Sahlgrenska Academy, University of Gothenburg, Sweden. Briefly, samples were analysed on an QExactive HF mass spectrometer and acquired spectra were processed using MaxQuant software suite V1.5.1.0 (Cox and Mann, 2008; Tyanova, Temu and Cox, 2016) integrated with an Andromeda (Cox et al. 2011) search engine. Database search was performed against a target-decoy database of *B. subtilis* (NCBI Assembly accession: ASM205596v1; GCA_002055965.1) containing 4,333 protein entries. Finally, iBAQ values for 2,915 proteins at 10% false discovery rate were obtained. Details about biofilm cultivating, RNA extraction and protein digestion can be found in Futo *et al.* (2021).

3.2. Transcriptome data analyses

Ribosomal RNA was removed from the total RNA samples and RNA-seq libraries were prepared using the Illumina TruSeq RNA Sample Preparation v2 Kit (Illumina). The RNA sequencing was performed bi-directionally on the Illumina NextSeq 500 platform at the EMBL Genomics Core Facility (Heidelberg, Germany), generating ~450 million reads per run. The sequence quality and read coverage were checked using FastQC V0.11.7 (Andrews, 2010) with satisfactory outcome for each of the samples. In total 1,448,793,058 paired-end sequences

(75bp) were mapped onto the *B. subtilis* reference genome (NCBI Assembly accession: ASM205596v1; GCA_002055965.1) using BBMap V37.66 (Bushnell, 2014) with an average of 93.46% mapped reads per sample (Appendix 1). We mapped in average 49 million reads per replicate with rather low variation between the samples (Appendix 1). The mapping was performed using the standard settings with the option of trimming the read names after the first whitespace enabled. We used the *SAMtools* package V1.6 (H. Li *et al.*, 2009) to generate, sort and index BAM files for downstream data analysis. The subsequent RNAseq data processing was performed in R V3.4.2 (R Development Core Team, 2008) using custom-made scripts. Briefly, mapped reads were quantified per each *B. subtilis* open reading frame using the R *rsamtools* package V1.30.0 (Morgan *et al.*, 2017) and raw counts for 4,515 open reading frames were retrieved using the *GenomicAlignments* R package V1.14.2 (Lawrence *et al.*, 2013). Expression similarity across timepoints and replicates was assessed using principal component analysis (PCA; Figure 1c) implemented in the R package *DESeq2* V1.18.1 (Love, Huber and Anders, 2014) and visualized using custom-made scripts based on the R package *ggplot2* V3.1.0. (Wickham, 2016) (Figure 1b).

Out of 4,333 protein coding genes with mapped reads, we analysed 4,316 which passed the phylostratigraphic procedure (Table 1). First, raw counts of these 4,316 protein coding genes were normalized by calculating the fraction of transcripts (τ) (B. Li *et al.*, 2009). After this normalization step, we resolved replicates by calculating their median, whereby replicates that had zero raw values were omitted, and such values were used for calculating transcriptome evolutionary indices (Figure 4a and c, 5a and c, 6, Appendix 4). We discarded the genes which had zero expression values in more than one stage in preparation of the transcriptome dataset for RNA expression profile correlations, clustering and visualization (Figure 7, 8-10, Appendix 2 and 3), thus reducing the dataset to 4,296 genes. If a gene had only one stage with a zero-expression value, the zero-value was imputed by interpolating the mean of the two neighbouring stages (2 genes in total). In the case a zero-expression value was present in the first or the last stage of the biofilm ontogeny, the value of the only neighbour was directly assigned to it (134 genes in total). This procedure was chosen to avoid erratic patterns in the visualization and clustering of mRNA expression profiles. To bring the gene expression profiles to the same scale, we performed the normalization to median and $\log_2$ transformation for every gene. This per-gene normalized expression values (standardized expressions) were used for visualization (Figure 8-10, Appendix 2 and 3) using the R *ggplot2* (Wickham, 2016) package V3.1.0 and clustered (Appendix 2 and 3) with DP_GP_cluster (McDowell *et al.*, 2018) with the maximum Gibbs sampling iterations set to 500. For transcription regulator and sigma factor

expression profiles (Figure 8a and b), we selected genes which are regulating ≥ 10 operons based on the DBTBS database (Sierro *et al.*, 2008). Biofilm-important genes and cell-cell signalling genes used for profile visualization (Figure 8c and f) were selected following relevant reviews (Vlamakis *et al.*, 2013; van Gestel, Vlamakis and Kolter, 2015; Kalamara *et al.*, 2018). Protein phosphatase and protein kinase genes were selected for profile visualization (Figure 8d and e) following the SubtiWiki database annotations (Zhu and Stülke, 2018). The statistical significance of difference between average standardized expressions shown in Figure 8 was assessed by repeated measures ANOVA. To determine the similarity of transcriptomes across stages of biofilm ontogeny, we calculated Pearson's correlation coefficients (R) in all-against-all manner and visualized them in a heat map (Figure 1b). Pairwise differential gene expression between individual stages of biofilm ontogeny (Figure 3) was estimated using a procedure implemented in the *DESeq2* (Love, Huber and Anders, 2014) V1.18.1 package. Using the likelihood ratio test implemented in the same package, the overall differential expression for every gene across all stages of biofilm ontogeny was tested (Figure 2).

3.3. Proteome data analyses

Out of 2,915 quantified proteins, we further analysed 2,907 which passed the phylostratigraphic procedure. First, we calculated the partial concentrations by dividing every iBAQ value by the sum of all iBAQ values in the sample. After this normalization step, we resolved replicates by calculating their median whereby replicates that had zero iBAQ values were omitted. This yielded normalized protein expression values that were used for calculating proteome evolutionary indices (Figure 4c and d, 5c and d). In preparing the proteome dataset for protein expression profile correlation, clustering and visualization (Figure 7, Appendix 3), genes which had zero-expression values in more than one stage were discarded, thus reducing the dataset to 2,543 proteins. If a protein had only one stage with a zero-expression value, it was interpolated by taking the mean of the two neighbouring stages (134 proteins in total). In the case a zero-expression value was present in the first or the last stage of the biofilm ontogeny, it was directly assigned the value of the only neighbour (355 proteins in total). To bring the protein expression profiles to the same scale, for every protein the normalization to median and $\log_2$ transformation was performed and yielded 2,543 per-protein normalized expression values (standardized expressions). These values were clustered (Appendix 3) with DP_GP_cluster (McDowell *et al.*, 2018) with the maximum Gibbs sampling iterations set to 500 and visualized in R *ggplot2*

package (Wickham, 2016) V3.1.0. To assess correlations between standardized transcriptome and proteome expression values, we calculated the Pearson's correlation coefficient (R) on the matching 2,543 genes/proteins (Figure 7f and g). Expression similarity across timepoints and replicates for 2,910 proteins was assessed using PCA in R V3.4.2 *stats* package (R Development Core Team, 2008). The proteome PCA plot was visualized (Figure 1d) using the R package *ggplot2* V3.1.0 (Wickham, 2016), with $\log_2$ transformed iBAQ values previously increased by 1.

3.4. Enrichment analysis

Due to the poor gene annotations for focal *B. subtilis* strain, we transferred annotations from *Bacillus subtilis* subsp. *subtilis* str. 168 (NCBI Assembly accession: ASM904v1; GCA_000009045.1) by establishing orthologs between the two strains. This was performed by calculating their reciprocal best hit using the blastp algorithm V2.7.1+ (Altschul *et al.*, 1990) with 10^{-5} e-value cut-off. Functional annotations for the *B. subtilis* 168 strain were retrieved from the SubtiWiki database (Zhu and Stülke, 2018) (accessed October 23, 2019). The list of orthologous genes and their affiliation to SubtiWiki functional annotations can be seen in Futo *et al.* (2021). Functional enrichment of individual biofilm timepoints was estimated using the one-tailed hypergeometric test. For each timepoint, we tested genes that had the expression in that timepoint 0.5 times ($\log_2$ scale) above the median of their overall expression profile for functional enrichment (Figure 11, Appendix 7-10). *P* values were adjusted for multiple comparisons using the Yekutieli and Benjamini procedure (Yekutieli and Benjamini, 1999).

3.5. Evolutionary measures

Genomic phylostratigraphy is a method that traces the evolutionary origin of genes by similarity searches in genomes representing the phylogeny of an organism of interest. This way it establishes a phylogenetic scale and for every gene in a genome of an organism of interest assigns a phylogenetic rank (Domazet-Lošo, Brajković and Tautz, 2007; Domazet-Lošo and Tautz, 2010). The phylostratigraphic procedure for *B. subtilis* was performed as previously described (Domazet-Lošo, Brajković and Tautz, 2007; Domazet-Lošo and Tautz, 2010).

Briefly, a consensus phylogeny that covers divergence from the last common ancestor of cellular organisms to the *B. subtilis* as a focal organism was previously constructed by the Domazet-Lošo group (Laboratory of Evolutionary Genetics, Department of Molecular Biology, Ruđer Bošković Institute). A full set of protein sequences for 926 terminal taxa were retrieved from online databases. Only the longest splicing variant per gene for eukaryotic organisms was left, and taxon tags were added to the sequence headers of all sequences. The details about constructing a phylogeny and preparing sequences for sequence similarity searches can be found in Futo *et al.* (2021). To construct the phylostratigraphy map of *B. subtilis*, we compared 4,333 *B. subtilis* proteins to the protein sequence database by blastp algorithm (Altschul *et al.*, 1990) V2.7.1+ with the 10^{-3} e-value threshold. After discarding all protein sequences which did not return their own sequence as a match, 4,317 protein sequences left in the sample. These 4,317 protein sequences were then mapped on the 12 internodes (phylostrata) of the consensus phylogeny using a custom-made pipeline (Table 1). Using expression values for 4,316 protein coding genes, we calculated the transcriptome age index (TAI; Figure 4a); *i.e.* the weighted mean of phylogenetic ranks (phylostrata) for each ontogenetic stage. To test for the robustness of the phylostratigraphic procedure and the TAI profile, several phylostratigraphic maps with different e-value cut-offs (10, 1, 10^{-1}, 10^{-2}, 10^{-5}, 10^{-10}, 10^{-15}, 10^{-20} and 10^{-30}) were constructed (Figure 6, Appendix5) using the blastp algorithm (Altschul *et al.*, 1990) V2.7.1+.

To analyse the proteome data in similar fashion, we introduced the proteome age index (PAI; Figure 4c):

$$PAI = \frac{\sum_{i=1}^{n} ps_i q_i}{\sum_{i=1}^{n} q_i}$$

where ps_i is an integer which represents the phylostratum of the protein i, q_i is iBAQ value of the protein i that acts as weight factor and n is the total number of proteins analysed. In total, 2,907 proteins were used for PAI calculation. To estimate evolutionary divergence rates of *B. subtilis* proteins, we found 3,094 orthologs in *Bacillus licheniformis* str. DSM 13 (NCBI Assembly accession: ASM1164v1; GCA_000011645.1) by reciprocal best hits using blastp with 10^{-5} e-value threshold. The list of orthologous gene pairs between *B. subtilis* and *B. licheniformis* can be seen in Futo *et al.* (2021). Of the available *Bacillus* species, *B. licheniformis* provided the best balance between the number of detected orthologues and the evolutionary distance of species pair for calculating evolutionary rates. Orthologous pairs between *B. subtilis* and *B. licheniformis* were globally aligned and codon alignments were constructed in pal2nal (Suyama, Torrents and Bork, 2006). We calculated the non-synonymous substitution rate (dN) and the synonymous substitution rate (dS) using the Comeron's method

(Comeron, 1995). The whole procedure of obtaining dN and dS was performed in the R package *orthologr* V0.3.0.9000 (Drost *et al.*, 2015). Using dN values of 3,091 genes, we calculated the transcriptome nonsynonymous divergence index (TdNI; Figure 4b), *i.e.* the weighted arithmetic mean of nonsynonymous gene divergence:

$$TdNI = \frac{\sum_{i=1}^{n} dN_i e_i}{\sum_{i=1}^{n} e_i}$$

where dN_i is a real number which represents the nonsynonymous divergence of gene *i*, e_i is the transcriptome expression value of the gene *i* that acts as weight factor and *n* is the total number of genes analysed. Using dS values of 2,212 genes, we calculated transcriptome synonymous divergence index (TdSI; Figure 5a), *i.e.* the weighted arithmetic mean of synonymous gene divergence:

$$TdSI = \frac{\sum_{i=1}^{n} dS_i e_i}{\sum_{i=1}^{n} e_i}$$

where dS_i is a real number which represents the synonymous divergence of gene *i*, e_i is the transcriptome expression value of the gene *i* that acts as weight factor and *n* is the total number of genes analysed. To analyse proteome data in similar fashion, we introduced the PdNI and PdSI. Using dN values of 2,329 proteins, we calculated the proteome nonsynonymous divergence index (PdNI; Figure 4d), *i.e.* the weighted arithmetic mean of nonsynonymous divergence:

$$PdNI = \frac{\sum_{i=1}^{n} dN_i q_i}{\sum_{i=1}^{n} q_i}$$

where dN_i is a real number which represents the nonsynonymous divergence of protein *i*, q_i is the normalized protein expression value of the protein *i* that acts as weight factor and *n* is the total number of proteins analysed. Using expression values of 1,755 proteins, we calculated proteome synonymous divergence index (PdSI; Figure 5b), *i.e.* the weighted arithmetic mean of synonymous divergence:

$$PdSI = \frac{\sum_{i=1}^{n} dS_i q_i}{\sum_{i=1}^{n} q_i}$$

where dS_i represents the synonymous divergence value of protein *i*, q_i is the normalized protein expression value of protein *i* that acts as weight factor and *n* is the total number of proteins analysed. Using 4,316 transcriptome expression values and "measure independent of length and composition" (MILC) (Supek and Vlahoviček, 2005) values, we calculated transcriptome codon usage bias index (TCBI; Figure 5c), *i.e.* the weighted arithmetic mean of codon usage bias:

$$TCBI = \frac{\sum_{i=1}^{n} MILC_i e_i}{\sum_{i=1}^{n} e_i}$$

where *MILC* is a real number which represents the codon usage bias of gene *i*, e_i is the transcriptome expression value of the gene *i* that acts as weight factor and *n* is the total number of genes analysed. Using 2,907 protein expression and MILC values, we calculated proteome codon usage bias index (PCBI; Figure 5d), *i.e.* the weighted arithmetic mean of codon usage bias:

$$PCBI = \frac{\sum_{i=1}^{n} MILC_i q_i}{\sum_{i=1}^{n} q_i}$$

where *MILC* is a real number which represents the codon usage bias of protein *i*, q_i is the normalized protein expression value of protein *i* that acts as weight factor and *n* is the total number of proteins analysed. MILC values were obtained from R package coRdon (Elek, Kuzman and Vlahoviček, 2020) V1.3.0, with respect to codon usage bias of ribosomal genes. The whole procedure of obtaining TdNI, TdSI, PdNI and PdSI was made in R package *orthologr* V0.3.0.9000. The statistical analysis for TAI, PAI, TdNI, TdSI, PdNI, PdSI, TCBI, PCBI was calculated using the R package *myTAI* V0.9.0 (Drost *et al.*, 2018).

4. RESULTS AND DISCUSSION

4.1. Biofilm growth is a stage-organised process

To measure transcriptome expression levels during *B. subtilis* biofilm formation, we sampled eleven timepoints covering a full span of biofilm ontogeny from its inoculation, until two months of age (Figure 1a). The transcriptome expression values were recovered for 4,316 (96%) *B. subtilis* genes by RNAseq, which revealed three distinct periods of biofilm ontogeny: early (6H-1D), mid (3D-7D) and late period (1M-2M), linked by two transition stages at 2D and 14D (Figure 1b). Biofilm transcriptomes also showed poor correlation to the liquid culture (LC) used for inoculation of biofilms, indicating that biofilm makes a distinct part of the *B. subtilis* life cycle (Figure 1b). Poor correlation to the LC was also shown on the principal component analysis (PCA), whereas timepoints after inoculation show a time-resolved profile (Figure 1c and d). The observed differences in gene expression during *B. subtilis* biofilm growth are in concordance with previous studies and confirm that biofilm formation is genetically dynamic mechanism (Vlamakis *et al.*, 2008).

Considering all ontogeny timepoints, 4,263 (99%) genes were differentially expressed (Figure 2). This number stayed similar (4,190 genes, 97%) when we looked only at biofilm growth in a narrow sense (6H-14D, Figure 2) by excluding the starting liquid culture (LC) and late-period timepoints (1M-2M) that show biofilm growth arrest. When we retained only genes with two-fold or higher expression change, the numbers still remained high: 2,546 genes (59%) in biofilm growth in a narrow sense and 2,798 genes (65%) in biofilm growth in a broad sense (Figure 2). Pairwise comparisons between successive ontogeny timepoints uncovered that most genes (around 70%) change their transcription at biofilm inoculation (LC-6H), indicating that transition from a liquid culture to solid agar plates represents a dramatic shift in *B. subtilis* lifestyle (Figure 3a). The most dynamic step during biofilm growth is transition at 1D-2D, regardless which fold change criteria is applied (Figure 3). Transcription at 1D-2D changes in approximately 5-30% of genes, depending on the fold change criteria (Figure 3). Other memorable gene expression changes are at 7D-14D and 14D-1M, although they are less visible at greater fold changes (Figure 3).

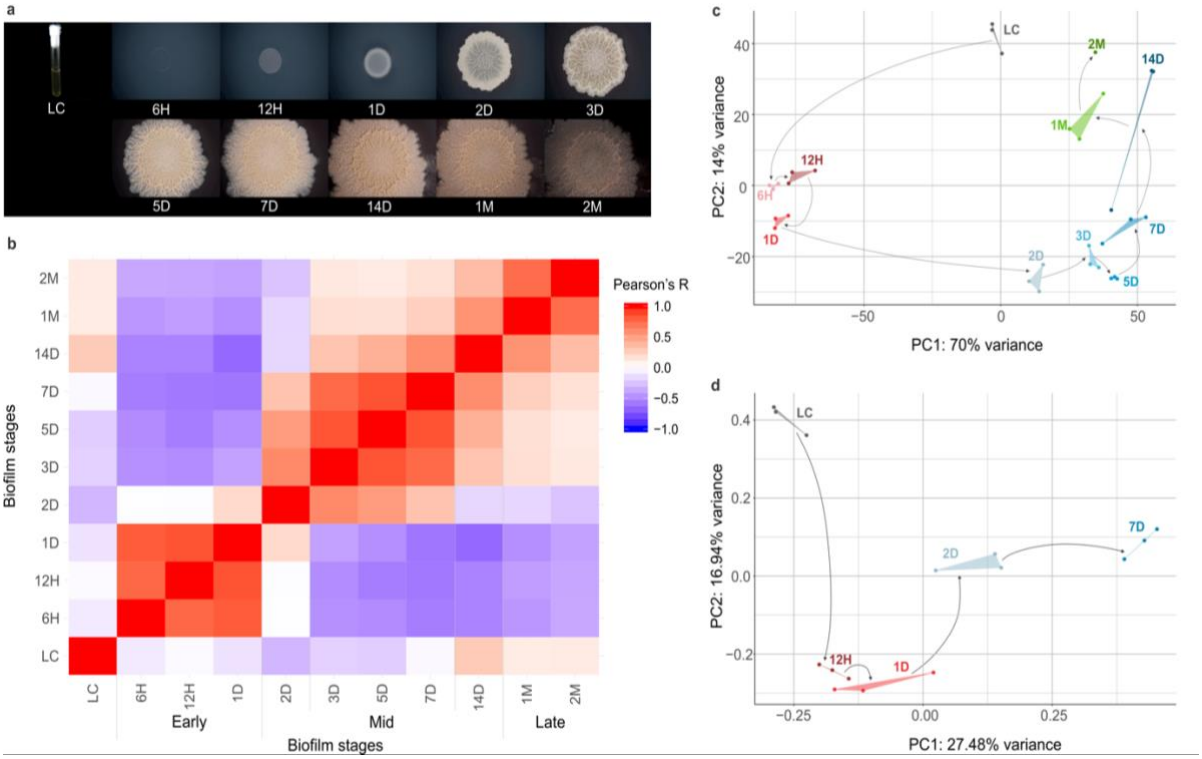


Figure 1. ***Bacillus subtilis*** **biofilm growth is a highly regulated and punctuated process. a)** Gross morphology of *B. subtilis* biofilms on solid agar plates at 6 hours (6H), 12 hours (12H), 1 day (1D), 2 days (2D), 3 days (3D), 5 days (5D), 7 days (7D), 14 days (14D), 1 month (1M) and 2 months (2M) after inoculation with liquid culture (LC). **b)** Pearson's correlation coefficients between timepoints of biofilm ontogeny in all-against-all comparison. Early (6H-1D), mid (3D-7D) and late (1M-2M) periods, together with transition stages at 2D and 14D, are marked. PCA of **c)** transcriptome and **d)** proteome data (see Material and Methodology). Biofilm growth timepoints (LC, 6H, 12H, 1D, 2D, 3D, 5D, 7D, 14D, 1M and 2M) are shown in different colours, where grey represents the liquid culture (LC), different shades of red early (6H-1D), blue mid (2D-14D) and green late (1M-2M) biofilm period. Replicates are in the same colour and connected with lines. Black arrows correspond to the experimental timeline of biofilm growth that starts with LC and ends at 2M. Figure modified from Futo *et al.* (2021).

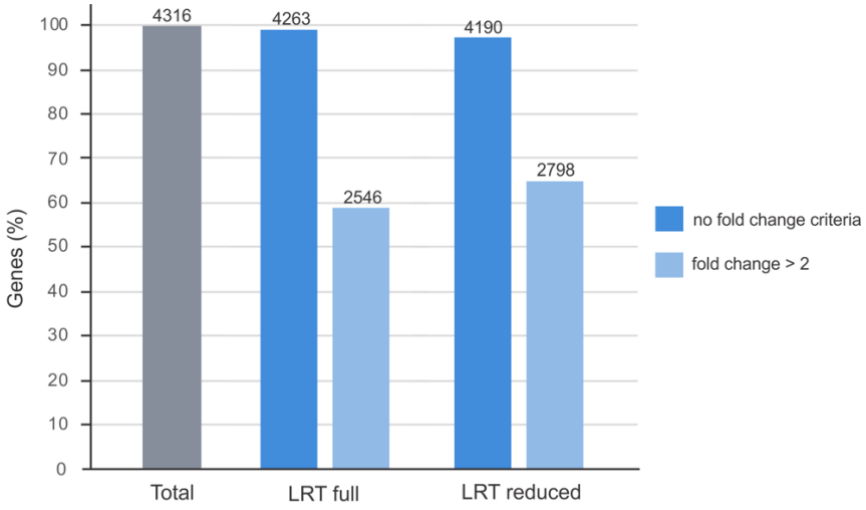


Figure 2. Differential expression through overall *B. subtilis* ontogeny shows dynamic regulation of transcription. The numbers above bars indicate the total number of genes recovered by RNAseq (**Total**), number of differentially expressed genes considering all ontogeny timepoints (**LRT full**), and number of differentially expressed genes considering ontogeny timepoints 6H – 14D (**LRT reduced**), with no fold change criteria, and genes with 2-fold or higher expression change. Differentially expressed genes were determined by likelihood ratio test (LRT) implemented in DESeq2, with *p* value (adjusted for multiple comparisons) cut-off set at 0.05.

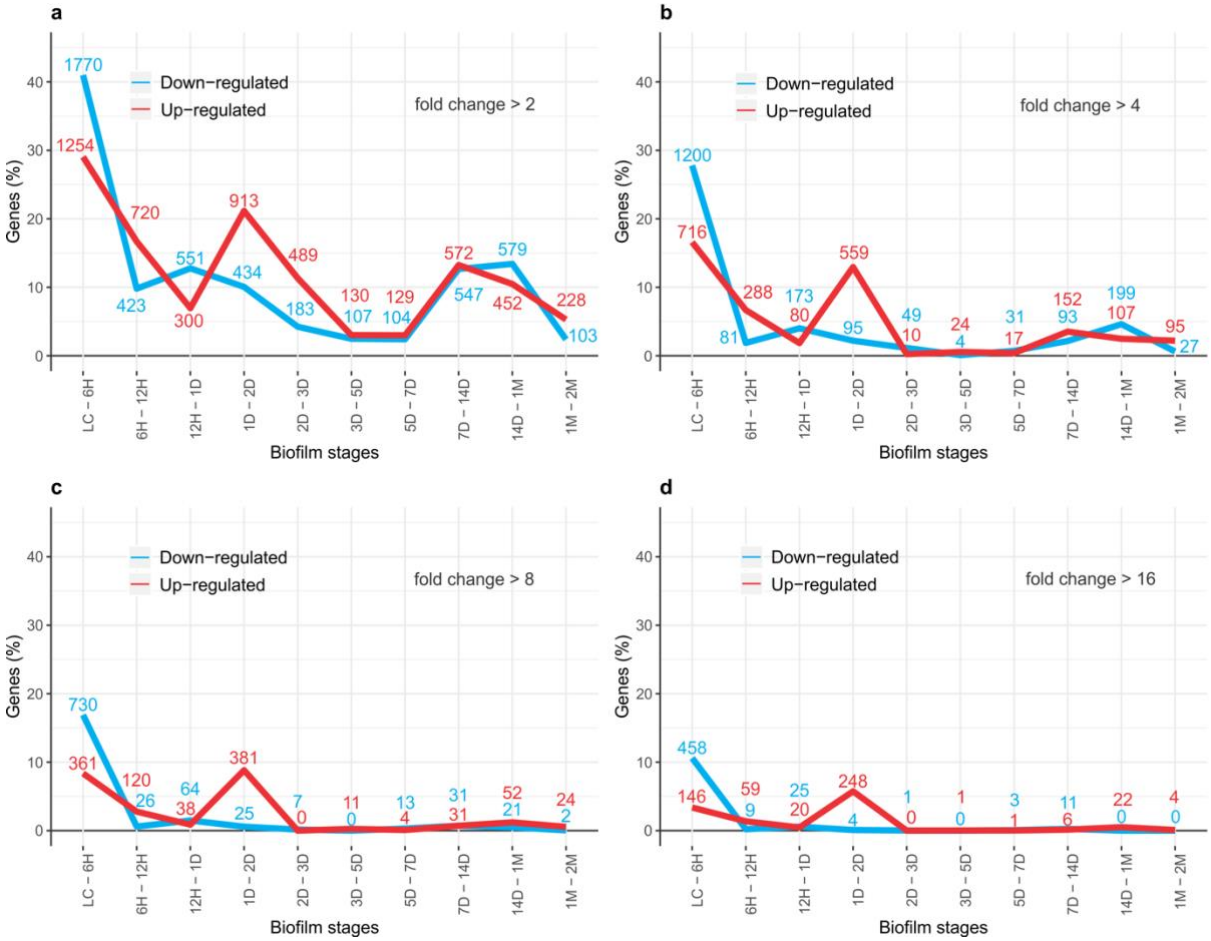


Figure 3. Differential expression in pairwise comparisons between successive biofilm timepoints. Bursts of differentially transcribed genes are visible after inoculation (LC-6H) and at 1D-2D, 7D-14D and 14D-1M transitions. The numbers above lines indicate the number of differentially expressed genes. Down-regulated genes are in blue, and up-regulated are in red. **a)** fold change cut-off > 2; **b)** fold change cut-off > 4; **c)** fold change cut-off > 8; **d)** fold-change cut-off > 16. Differentially expressed genes were determined by DESeq2 with *p* value (adjusted for multiple comparisons) cut-off set at 0.05. Figure obtained from Futo *et al.* (2021).

The changes in gene expression related to evolutionary age of genes during biofilm formation are comparable to those observed during animal ontogeny (Domazet-Lošo and Tautz, 2010). Differential gene expression (Figure 2, Figure 3), correlation between timepoints (Figure 1b), PCA of biofilm ontogeny (Figure 1c) and transcriptome clustering analysis (Appendix 2) support the idea that *B. subtilis* biofilm growth is intermittent with surges of transcriptional changes that define discrete ontogeny phases; the early, mid, and late period. This is similar to animal embryos, which show punctuated development both on morphological and transcriptomic level (Domazet-Lošo and Tautz, 2010; Kalinka *et al.*, 2010; Levin *et al.*, 2012; Yanai, 2018). The finding that *B. subtilis* biofilm growth is not a continuous process is additionally corroborated with PCA and clustering analysis of proteome (Figure 1d and

Appendix 3), although transcript and protein levels are known to poorly correlate (Liu, Beyer and Aebersold, 2016).

4.2. Evolutionary expression measures show a recapitulation pattern

To assess whether biofilm growth has some evolutionary directionality, or if there is no correlation between ontogeny and phylogeny, we linked transcriptome expression values to evolutionary gene age estimates (Table 1) to obtain the transcriptome age index (TAI); a cumulative measure that gives the overall evolutionary age of an expressed mRNA pool (Domazet-Lošo and Tautz, 2010; Quint *et al.*, 2012; Cheng *et al.*, 2015). If one assumes that expression patterns across biofilm ontogeny are independent of evolutionary age of genes, then the TAI profile should show a trend close to a flat line; that is, TAI and ontogeny should not correlate. We found a recapitulation pattern in *B. subtilis* biofilms, where early timepoints of biofilm growth express evolutionary older transcriptomes compared to mid and late timepoints that exhibit increasingly younger transcriptomes (Figure 4a). We also examined how the TAI profile relates to the evolutionary age of genes (phylostrata - ps) and found that recapitulation pattern is significant already from the origin of Firmicutes (Appendix 4), reflecting its rather deep roots in the bacterial phylogeny. It is important to note that in this context, the recapitulation term refers only to the transcriptional activation of genes along ontogeny which recapitulates the macroevolutionary sequence of gene emergence (Domazet-Lošo and Tautz, 2010). This term has remained useful in discussing development, although Haeckel's idea that ontogeny recapitulates phylogeny has been previously abandoned (Abzhanov, 2013).

Besides looking at the phylogeny-ontogeny correlation based on the emergence of founder genes through phylostratigraphy and TAI (Domazet-Lošo, Brajković and Tautz, 2007; Domazet-Lošo and Tautz, 2010), we also analysed the dataset by looking at more recent evolutionary history via estimating evolutionary divergence rates of coding sequences (Quint *et al.*, 2012). Quint *et al.* (2012) were measuring evolutionary divergence through the ratio of synonymous (dS) and nonsynonymous mutations (dN), because it is generally assumed that nonsynonymous substitution rates reflect selective pressure, and synonymous substitution rates demonstrate an estimate of neutral evolution in coding sequences. However, *B. subtilis* displays a strong codon usage bias (Sharp, 2005), so dS sites cannot be considered neutral with selection acting on them. To account for this, we looked at substitution rates separately by devising transcriptome nonsynonymous (TdNI) and synonymous (TdSI) divergence indices (see

Material and Methodology). In *B. subtilis* - *B. licheniformis* comparison, TdNI showed a recapitulation pattern from 1D onwards, where genes conserved at nonsynonymous sites tend to be used early, while more divergent ones are used later during the biofilm ontogeny (Figure 4b). Comparably, TdSI displays more complex correlation which clearly resembles the pattern of the transcriptome codon bias index (TCBI), indicating dependence of synonymous substitution rates and codon usage bias (Figure 5a and c). Nevertheless, TdSI recapitulation profile is evident in mid-period biofilms (1D-14D), where genes with more divergent synonymous sites gradually increase in transcription from 1D to 14D (Figure 5a). Together, these divergence-ontogeny parallelisms in *B. subtilis* biofilms further support the recapitulative evolutionary pattern and show that it is provided by relatively recent evolutionary forces in mid-period biofilms.

To test if the phylogeny-ontogeny correspondence also exists at the proteome level, we quantified proteomes of LC, 12H, 1D, 2D and 7D stages, which show typical *B. subtilis* biofilm morphology and present the most dynamic part of *B. subtilis* biofilm development. We obtained protein expression values for 2,907 (67%) predicted proteins and used them to calculate the proteome age index (PAI); a cumulative measure analogous to TAI (see Material and Methodology), that gives an overall evolutionary age of a protein pool. The PAI profile also showed a significant recapitulation pattern, where evolutionary older proteins have higher expression early and younger ones later during biofilm ontogeny (Figure 4c). Similar to TdNI, TdSI, and TCBI for transcriptome, proteome nonsynonymous (PdNI; Figure 4d) and synonymous (PdSI; Figure 5b) divergence indices in *B. subtilis* - *B. licheniformis* comparison, as well as proteome codon bias index (PCBI; Figure 5d) revealed that recapitulation pattern also holds at shallower evolutionary levels (see Material and Methodology). Jointly, this demonstrates that phylogeny-ontogeny dependence, beside transcriptomes, is also visible in biofilm proteomes.

Table 1. Phylostratigraphy map of *Bacillus subtilis* subsp. *subtilis* NCIB 3610. The table shows the distribution of *B. subtilis* genes on the phylostratigraphy map and summary statistics with 1e-3 BLAST e-value cut-off. The abbreviation FACCAM (ps3) stands for Firmicutes, Actinobacteria, Chloroflexi, Cyanobacteria, Armatimonadates and Melainabacteria. Table obtained from Futo *et al.* (2021).

Phylostratum number	Phylostratum name	Number of genes	Percentage of genes
1	Cellular organism	2563	59.37%
2	Bacteria	597	13.83%
3	FACCAM	80	1.85%
4	Firmicutes	444	10.28%
5	Bacilli	24	0.56%
6	Bacillales	71	1.64%
7	Bacillaceae	145	3.36%
8	Bacillus	174	4.03%
9	Bacillus subtilis group	89	2.06%
10	Bacillus subtilis	32	0.74%
11	Bacillus subtilis subspecies subtilis	75	1.74%
12	Bacillus subtilis subspecies subtilis NCIB 3610	23	0.53%
	Total:	4317	100.00%

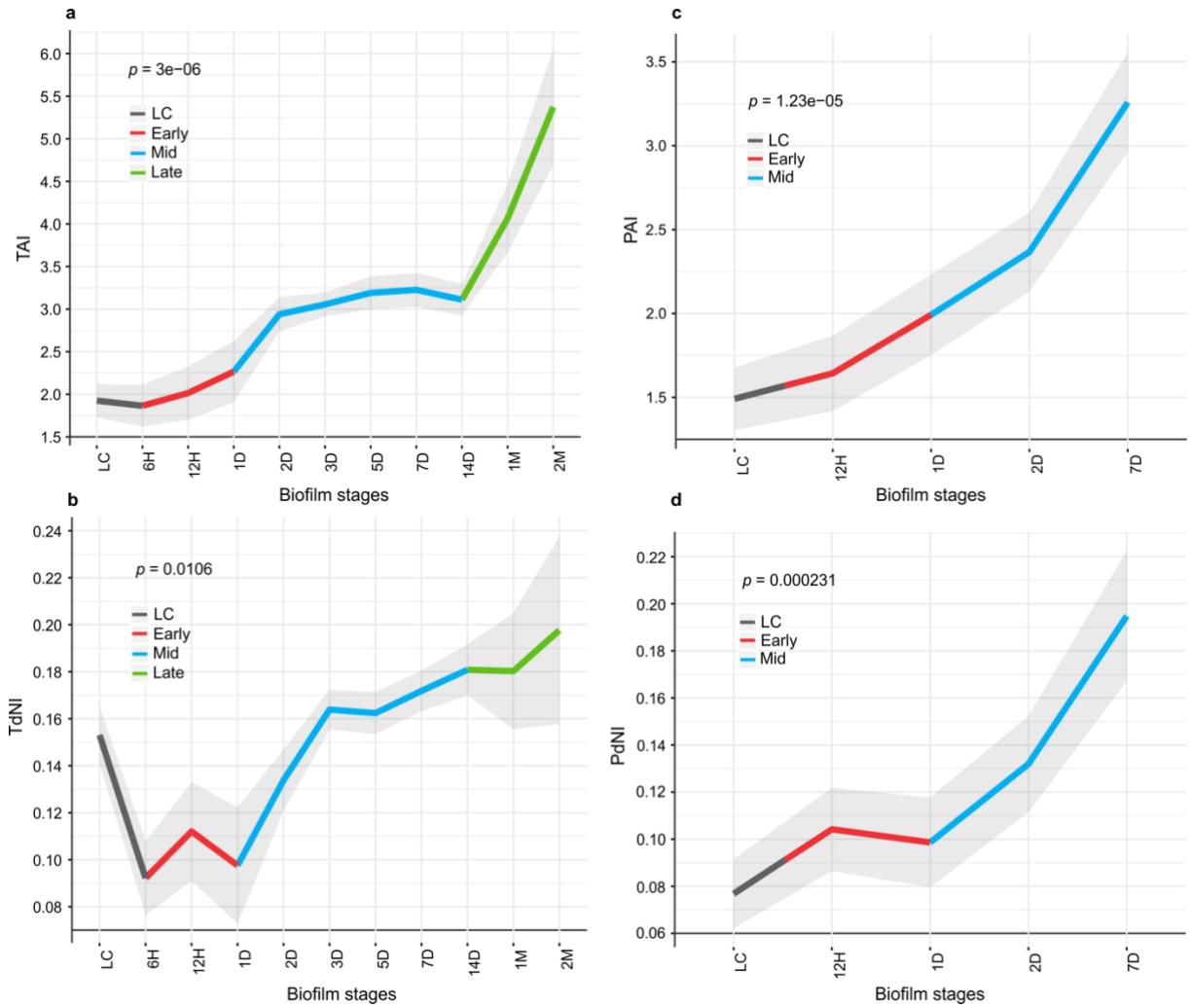


Figure 4. ***Bacillus subtilis*** **biofilm growth exhibits a phylogeny-ontogeny recapitulation pattern. a)** Transcriptome age index (TAI) and **b)** proteome age index (PAI) profiles of *B. subtilis* show that evolutionary older genes are used early in the biofilm ontogeny, while evolutionary younger genes are used later during biofilm ontogeny. **c)** Transcriptome nonsynonymous divergence index (TdNI) and **d)** proteome nonsynonymous divergence index (PdNI) profiles show that genes conserved at nonsynonymous sites are used early in the biofilm ontogeny, while more divergent ones are used later during biofilm ontogeny. Nonsynonymous divergence rates were estimated in *B. subtilis* – *B. licheniformis* comparison. Depicted *p* values are obtained by the flat line test and grey shaded areas represent ± one standard deviation estimated by permutation analysis (see Material and Methodology). Early (red), mid (blue) and late (green) periods of biofilm growth are colour coded. Figure modified from Futo *et al.* (2021).

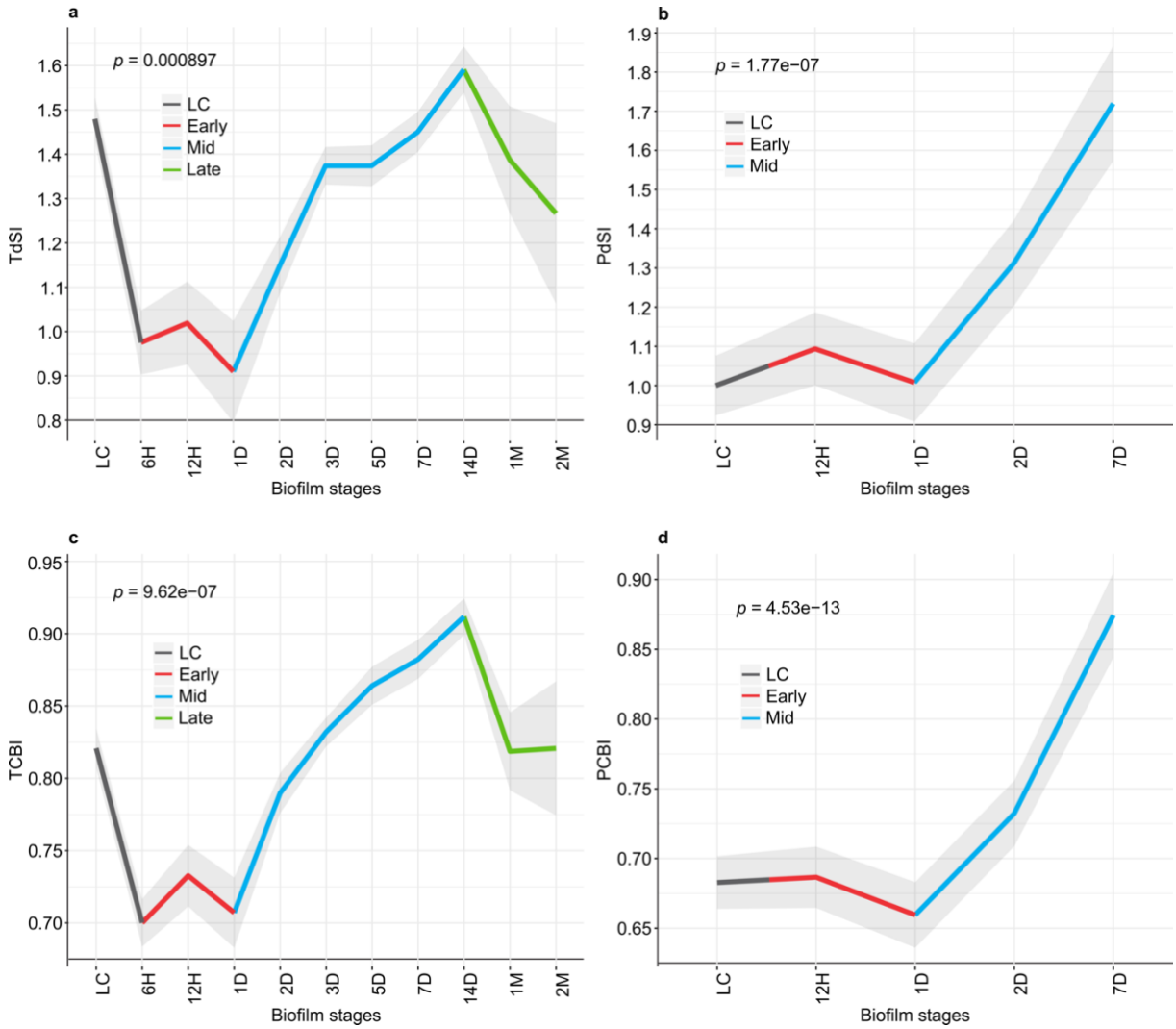


Figure 5. Synonymous divergence and codon usage bias are co-dependent. a) Transcriptome synonymous divergence index (TdSI) shows resemblance to **c)** transcriptome codon bias index (TCBI). **b)** Proteome synonymous divergence index (PdSI) shows resemblance to **d)** proteome codon bias index (PCBI). Depicted *p* values are obtained by the flat line test and grey shaded areas represent ± one standard deviation estimated by permutation analysis (see Material and Methodology). Early (red), mid (blue) and late (green) periods of biofilm growth are colour-coded. Figure obtained from Futo *et al.* (2021).

Phylostratigraphy based-tools were already successfully applied in detecting phylogeny-ontogeny correlations, not only in animals (Domazet-Lošo and Tautz, 2010), but also in plants (Quint *et al.*, 2012) and fungi (Cheng *et al.*, 2015). Moreover, several unrelated approaches independently also revealed the connection between phylogeny and ontogeny on molecular level in Metazoa (Kalinka *et al.*, 2010; Irie and Kuratani, 2011; Levin *et al.*, 2012, 2016), so phylostratigraphy can be considered as a reliable approach. Nevertheless, we evaluated the robustness of the TAI recapitulation pattern in *B. subtilis* by changing the BLAST e-value cut-off values in the broad range from 10 to 10^{-30} (Figure 6, Appendix 5). High e-value cut-offs inflate the false positive rates and push gene ages towards older phylostrata, while low e-value cut-offs boost false negative rates and classify genes into younger phylostrata (Appendix 5). Regardless of the shifts in the distribution of genes across phylostrata, the TAI recapitulation pattern remained stable and significant (Figure 6, Appendix 5), showing a definite macroevolutionary imprint in the *B. subtilis* biofilm ontogeny that is resistant to changes in BLAST e-value thresholds.

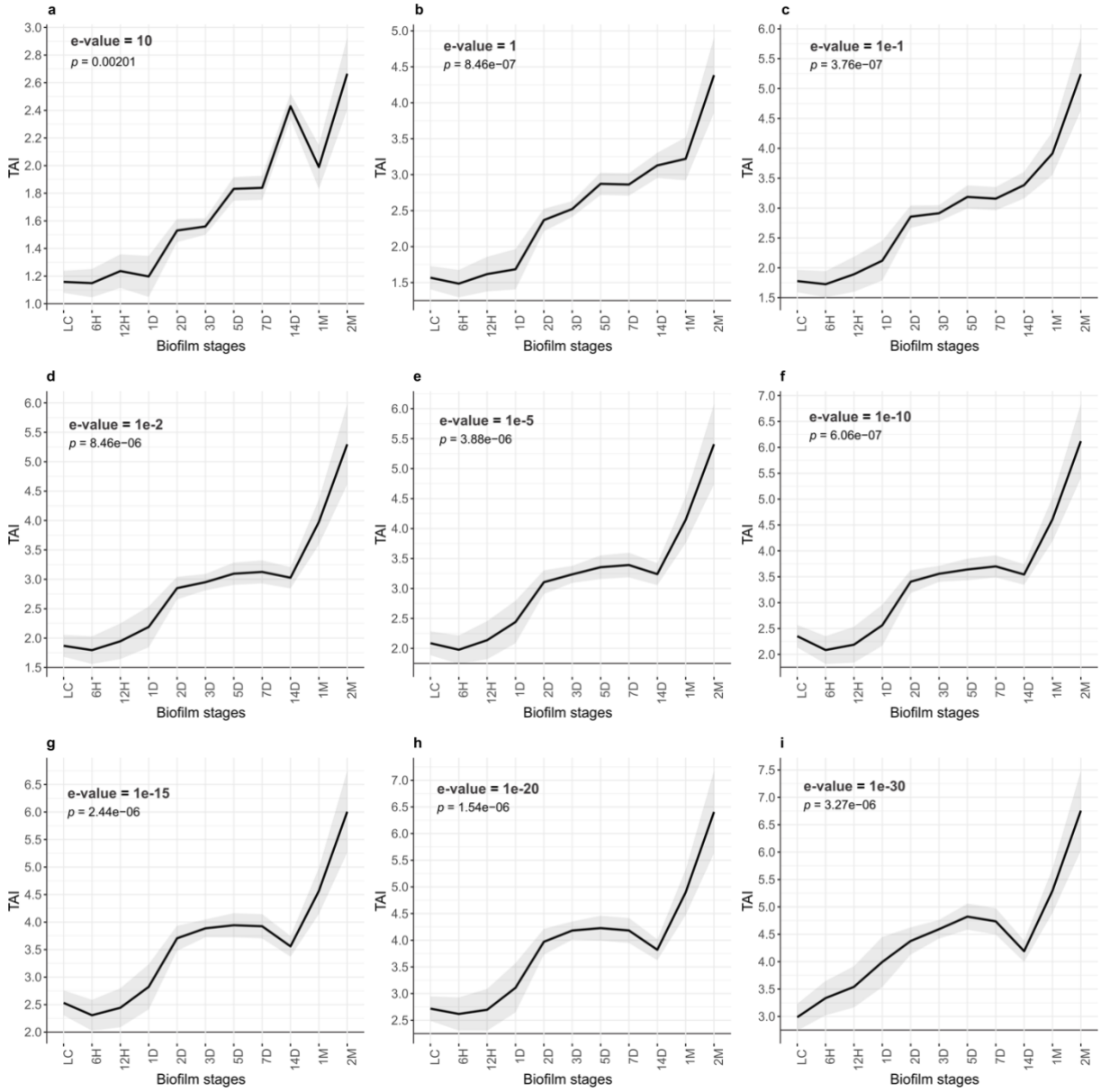


Figure 6. Transcriptome age index profiles of *B. subtilis* ontogeny for different BLAST e-value cut-offs. a) e-value = 10 (n = 4,322 genes); **b)** e-value = 1 (n = 4,318 genes); **c)** e-value = 10^{-1} (n = 4,318 genes); **d)** e-value = 10^{-2} (n = 4,317 genes); **e)** e-value = 10^{-5} (n = 4,315 genes); **f)** e-value = 10^{-10} (n = 4,305 genes); **g)** e-value = 10^{-15} (n = 4,291 genes); **h)** e-value = 10^{-20} (n = 4,274 genes); **i)** e-value = 10^{-30} (n = 4,170 genes). Depicted *p* values are obtained by the flat line test and grey shaded areas represent ± one standard deviation estimated by permutation analysis (see Material and Methodology**)**. Distribution of genes on phylostratigraphy maps with different BLAST e-value cut-off can be seen in Appendix 5. Figure obtained from Futo *et al.* (2021).

Besides being robust, the TAI profile of *B. subtilis* biofilm shows a correspondence between phylogeny and ontogeny that is similar to those of multicellular organisms (Domazet-Lošo and Tautz, 2010; Quint *et al.*, 2012; Cheng *et al.*, 2015). Although multicellularity evolved independently in these lineages (Niklas, 2014), it seems that it is governed by similar basic principles that include a macroevolutionary imprint.

Despite of the relatively poor correlation between evolutionary age and sequence divergence (Quint *et al.*, 2012) and methodological independence of phylostratigraphy and divergence-based tools, it is interesting that evolutionary indices based on these two methods show the same recapitulation pattern. Furthermore, transcriptome and proteome levels within timepoints (Figure 7a-e) and across ontogeny (Figure 7f and g) in *B. subtilis* biofilms show no correlation, but protein evolutionary indices again show recapitulation pattern as transcriptome evolutionary indices. Although transcriptome can be regarded as a first step in expressing a phenotype from a genotype, proteome is closer to the phenotype and thus is required to fully understand processes like development (Peshkin *et al.*, 2015). There are some gene expressions that correlate with protein expressions (Figure 7i), but most of them show no correlation (Figure 7h), confirming rather poor connection between transcript and protein levels (Peshkin *et al.*, 2015; Liu, Beyer and Aebersold, 2016).

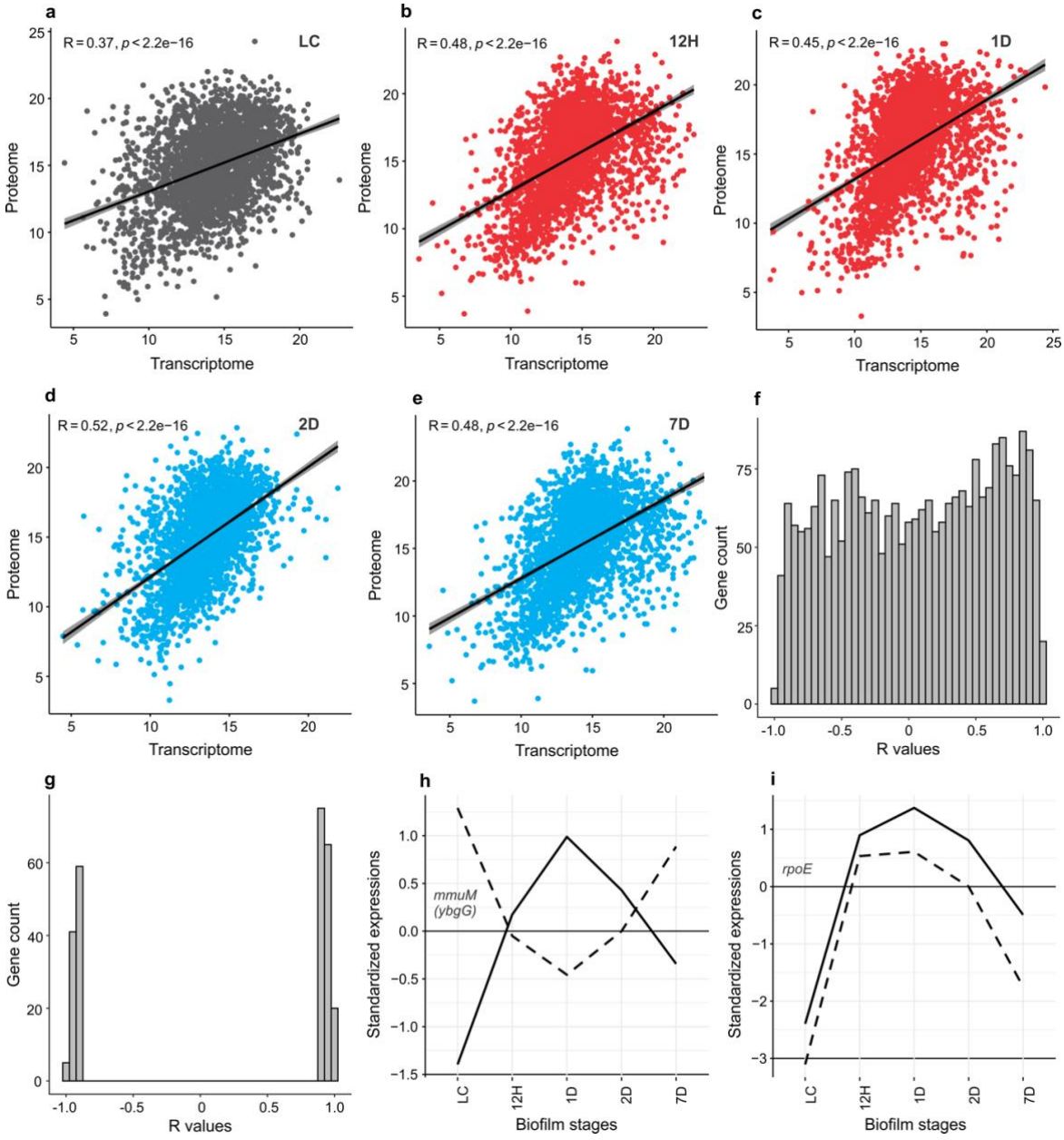


Figure 7. Transcriptome and proteome data show relatively low correlation. Correlation between transcriptome and proteome normalized expression values on -log$_2$ scale calculated for each biofilm timepoint separately. Pearson's correlation coefficients (R) with corresponding *p* values are shown. Genes that had zero expression values in either proteome or transcriptome were excluded. **a)** LC (n = 2,720 genes); **b)** 12H (n = 2,763 genes); **c)** 1D (n = 2,721 genes); **d)** 2D (n = 2,590 genes) and **e)** 7D (n = 2,318 genes). **f)** Distribution of Pearson's correlation coefficients (R) calculated for every gene between its transcriptome and proteome standardized expression profile across biofilm ontogeny (n = 2,543 genes). **g)** Distribution of only significant Pearson's correlation coefficients (R) ($p < 0.05$, n = 265 genes) from **f**. Significant negative correlation show 105 (4.1%) and significant positive 160 (6.3%) genes. **h)** An example of gene (*mmuM*; *ybgG B. subtilis* 168 strain name) that shows significant negative Pearson's correlation ($p = 0.009$, R = -0.96) between the transcriptome (dashed line) and the proteome (solid line) standardized expressions. **i)** An example of gene (*rpoE*) that shows significant positive Pearson's correlation ($p = 0.003$, R = 0.98) between the transcriptome (dashed line) and the proteome (solid line) standardized expressions. Figure obtained from Futo *et al.* (2021).

4.3. Multicellularity important genes dominate in mid-period biofilms

Transcription factors are one of the defining features of development in multicellular organisms, with dynamic expression throughout ontogeny (de Mendoza *et al.*, 2013). We found that *B. subtilis* transcription regulators cumulatively have the highest transcription in mid-period biofilms (2D to 7D), and that during this period majority of them are transcribed above the median of their overall expression profiles (Figure 8a, Appendix 6). This holds even if we narrow down the analysis to transcription initiators sigma factors only (Figure 8b, Appendix 6), and shows similarity with increased expression of transcription factors during embryo development (de Mendoza *et al.*, 2013). Communication between cells is another important aspect of coordinated multicellular function (Shapiro, 1998). Similar to transcription regulators, quorum sensing genes in *B. subtilis* biofilms peak in transcription at 3D (Figure 8c, Appendix 6), suggesting the most elaborate cell-cell communication at the timepoint when the biofilm gets the typical wrinkled morphology (Figure 1a). Protein phosphorylation is another important mechanism involved in many processes like differentiation, development, cell signalling or growth in animals and plants (Zhang and Liu, 2002; Xu and Zhang, 2015), and it also has an essential role in biofilm formation (López and Kolter, 2010; Vlamakis *et al.*, 2013; Kalamara *et al.*, 2018). Again, we found that protein kinases and phosphatases cumulatively have the highest transcription in mid-period biofilms (Figure 8d and e, Appendix 6), likely reflecting various types of cell differentiation in this growth phase (Vlamakis *et al.*, 2008). Furthermore, DNA methylation is known to impact many biological processes in eukaryotes, including development, and it has been revealed that DNA methylation status of enhancers during the animal phylotypic period controls the expression of developmental genes (Bogdanović *et al.*, 2016). Recently, it has been shown that DNA methylation has a function in regulating gene expression in bacteria also (Nye *et al.*, 2020). Interestingly, methyltransferase genes in *B. subtilis* biofilm show expression peak at the onset of biofilm formation (Figure 9), suggesting possible function of methylation in gene expression regulating biofilm formation.

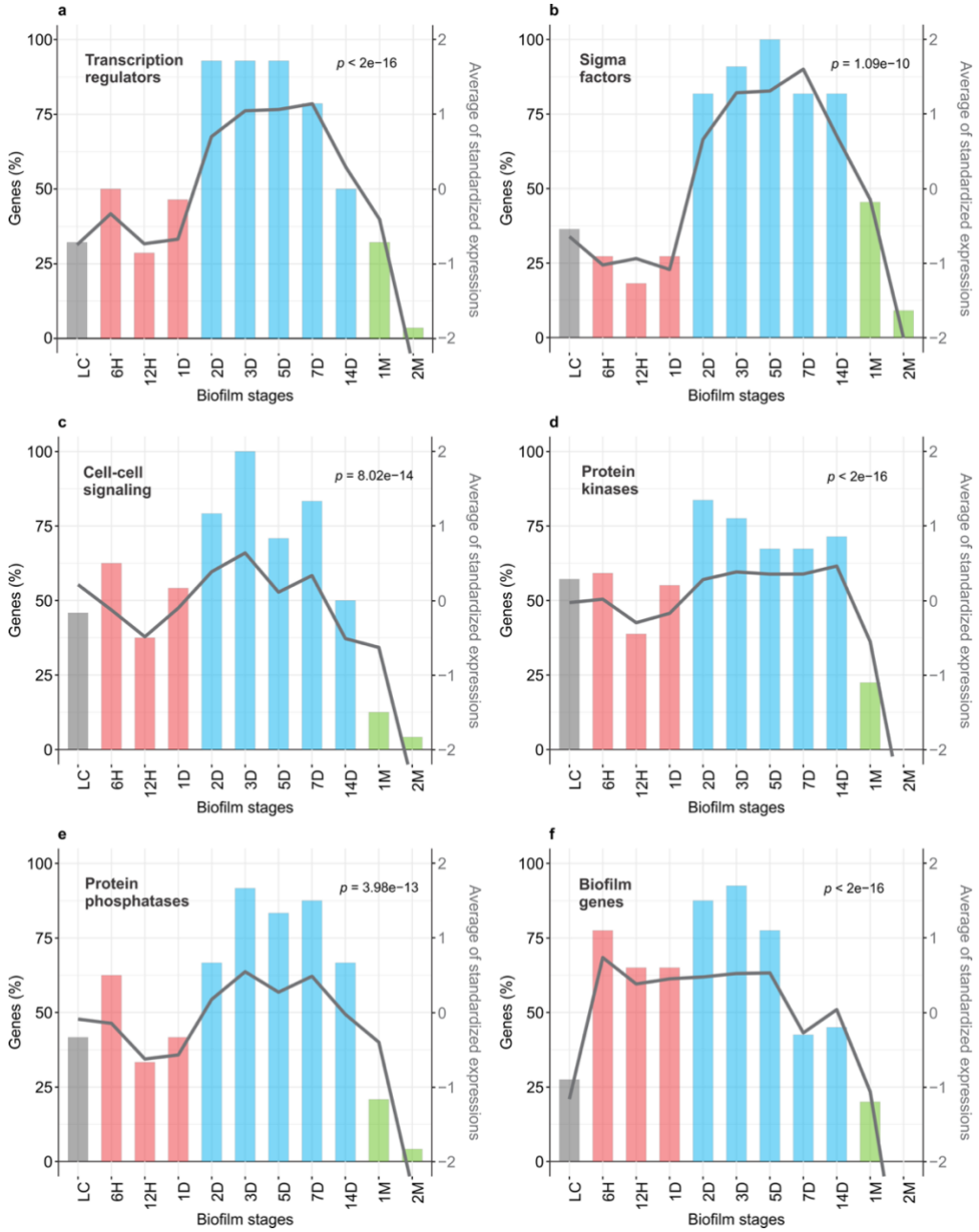


Figure 8. Multicellularity-important genes show cumulatively the strongest transcription in the mid-biofilm period. Left y-axis shows percentage of genes that are transcribed above the median of their overall transcription profile (histogram). Right y-axis shows the average standardized transcription values for all considered genes (line). Significance of the average expression profile is tested by repeated measures ANOVA and respective *p* values are shown. **a)** Transcription regulators that regulate ≥ 10 operons (see Material and Methodology, n = 28, $F(10, 270) = 17.33$); **b)** Sigma factors (n = 11, $F(10, 100) = 9.257$); **c)** Cell to cell signalling genes (n = 24, $F(10, 230) = 9.947$); **d)** Protein kinases (n = 49, $F(10, 480) = 41.71$); **e)** Protein phosphatases (n = 24, $F(10, 230) = 9.452$); **f)** Key biofilm genes (n = 40, $F(10, 390) = 30.74$). Colouring of bars in histograms follows biofilm growth periods: LC (grey), early (red), mid (blue), late (green). List of genes for categories "Transcription regulators", "Sigma factors", "Cell-cell signalling", "Protein kinases" and "Protein phosphatases" with their corresponding standardized expression values can be seen in Appendix 6. Figure obtained from Futo *et al.* (2021).

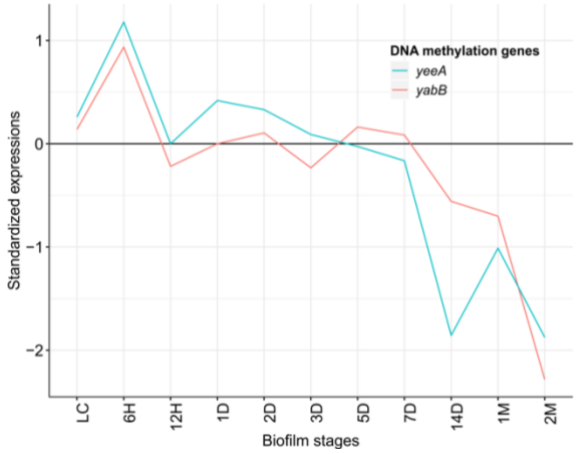


Figure 9. Methyltransferase genes show the highest expression at the onset of biofilm formation. Black horizontal line represents the median of standardized transcriptome expression values. Figure modified from Futo *et al.* (2021).

Further, we followed the expression of the key biofilm genes, and found that they are increasingly transcribed from the onset of biofilm formation (6H), maintain high values over early and mid-period (12H-14D), and progressively decline in late biofilms (1M-2M; Figure 8f). Individual profiles of the key biofilm genes further show their finer stratification and reflect their specific roles. For instance, extracellular matrix genes and genes responsible for positive control of biofilm formation show highest transcription in early biofilms (Figure 10a and f). Motility genes and genes involved in negative regulation of biofilm gradually decrease in expression as the biofilm forms (Figure 10b and c). Sporulation genes have the highest transcription in the mid-period biofilms (Figure 10h). Surfactin has bimodal distribution with peaks at 6H and 14D (Figure 10e), and protease production increases from 2D to 14D (Figure 10g).

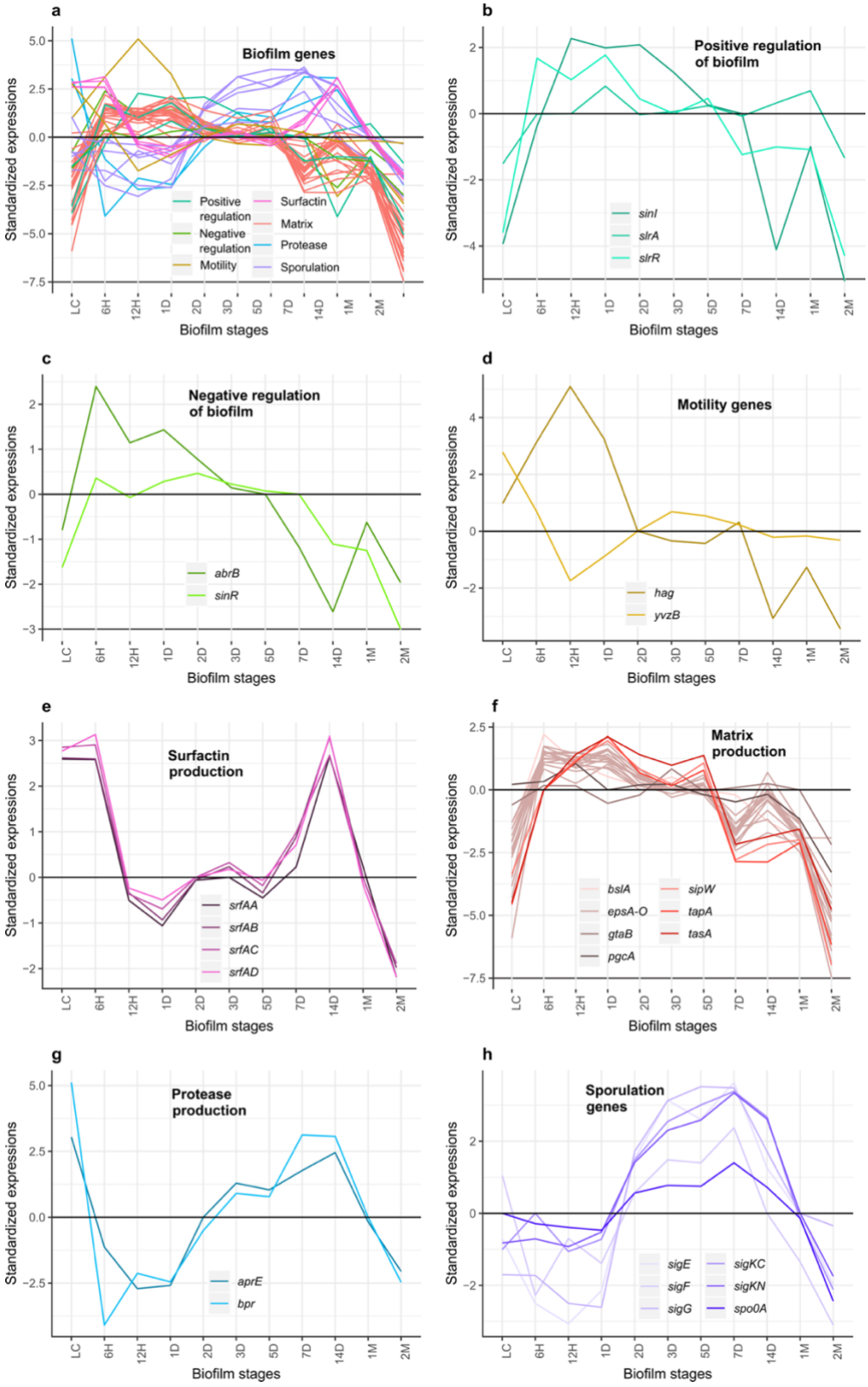


Figure 10. Standardized transcription profiles of key biofilm genes. a) All considered biofilm genes grouped in seven categories (n = 40); **b)** Positive regulators of biofilm growth (n = 3); **c)** Negative regulators of biofilm growth (n = 2); **d)** Biofilm important cell motility genes (n = 2); **e)** Surfactin genes (n = 4); **f)** Matrix genes (n = 21); **g)** Biofilm important proteases (n = 2); **h)** Sporulation genes (n = 6). Black horizontal line represents the median of standardized transcriptome expression values. Figure obtained from Futo *et al.* (2021).

These expression profiles are largely in accordance with previous knowledge on regulatory networks responsible for biofilm formation and other processes happening during that period (López and Kolter, 2010; Vlamakis *et al.*, 2013; Marlow *et al.*, 2014; van Gestel, Vlamakis and Kolter, 2015; Kalamara *et al.*, 2018). Expression profiles obtained in this study should also be beneficial in future work. Exploration of profiles of potential candidate genes, especially those with unknown function and with expression peaks in stages of interest, could lead to expanding the knowledge about regulatory networks in biofilm formation. Furthermore, the distribution of genes according to their evolutionary origin can additionally serve as a tool for assigning functions to genes, as predictive power of phylostratigraphy has already been demonstrated (Shi *et al.*, 2020).

The temporal dynamics of gene expressions in growing *B. subtilis* biofilms is similar to those in developing animals (Domazet-Lošo and Tautz, 2010; Kalinka *et al.*, 2010; Irie and Kuratani, 2011), and to other phyla exhibiting developmental processes (Quint *et al.*, 2012; Cheng *et al.*, 2015). With the uprise of molecular methods, research on animals has also established that many genes during development show tissue or cell specific expression (Tomancak *et al.*, 2007; Zeisel *et al.*, 2018; Cao *et al.*, 2019; Raj *et al.*, 2020). Single cell RNA sequencing also provides the ability to uncover the order in which the expression of different genes and regulatory networks happens during development and enlightens the basics of this process (Raj *et al.*, 2020). Although similar protocols are not yet established for studying biofilms, previous work has already shown that biofilm growth indeed shows spatial and temporal organization (Vlamakis *et al.*, 2008; Srinivasan *et al.*, 2018).

4.4. Biofilm growth has a stepwise functional architecture

The functional category enrichment analysis of biofilm timepoints reveals a stepwise architecture where every timepoint and biofilm stage express a specific group of functions (Figure 11, Appendix 7-10). Some illustrative examples of enriched functions include acquisition and biosynthesis of different macromolecules and components at the beginning of biofilm formation. Furthermore, swarming, motility, biofilm formation and phosphorylation are also enriched in the early biofilm. Phosphorylation continues to be enriched in mid-biofilm, along with sporulation, germination, and quorum sensing. Sporulation and germination are also important in the late biofilm, when different functions related to stress become enriched. The enrichment of genes that lack functional annotation probably reflects the incomplete knowledge

on the molecular mechanisms that govern early-to-mid biofilm transition. Statistical analysis of these genes on the phylostratigraphic map (Appendix 11) revealed that they preferentially originate from the ancestors of *B. subtilis* strains (ps10-ps12). This is similar to development in animals where taxonomically restricted genes are involved in generating the ontogenetic differentiation between taxa or morphological diversity in closely related species (Khalturin *et al.*, 2009; Tautz and Domazet-Lošo, 2011). When observed in total, functional enrichment patterns show that biofilm growth at the functional level has discrete hierarchical organization with even finer temporal grading compared to the pure transcription profiles. This punctuated and stage-like nature of biofilm growth is analogous to modular development in animals (Yanai, 2018).

Figure 11. Biofilm ontogeny is a punctuated process organized in functionally discrete stages. Enrichment analysis of SubtiWiki functional categories (ontology depth 3) in a respective biofilm growth timepoint for genes with transcription 0.5 times (log_2 scale) above the median of their overall transcription profile. Similar results are obtained for other transcription level cut-offs and SubtiWiki functional annotation ontology depths (see Appendix 7-10). Colouring follows biofilm growth periods: LC (grey), early (red), mid (blue), late (green). Functional enrichment is tested by one-tailed hypergeometric test and *p* values are adjusted for multiple testing (see Material and Methodology). Figure obtained from Futo *et al.* (2021).

Development can be defined as an emergence of organized structures from an initially simple group of cells, with pattern formation, polarity, changes in form, cell differentiation and growth as main processes that define it (Wolpert *et al.*, 2007). Cell differentiation with division of labour and changes in composition and number of different cell types, different spatio-temporal gene expression, and intercellular communication are well known properties of *B. subtilis* biofilms (Vlamakis *et al.*, 2008, 2013; López and Kolter, 2010; Kalamara *et al.*, 2018). On top of that, additional features that define multicellularity in *B. subtilis* include electrical communication and entering into proliferation phase that resembles cancer in animals (Prindle *et al.*, 2015; Hashuel and Ben-Yehuda, 2019). Our results implicate that stage-organized gene expression should be added to the growing list of properties that qualify *B. subtilis* biofilm as a multicellular organism.

5. CONCLUSION

Based on expression and quantification values of *B. subtilis* biofilm transcriptome and proteome, and on implementation of evolutionary measures, several conclusions can be derived from this study:

1. *B. subtilis* biofilm shows molecular phylogeny-ontogeny correspondence through recapitulation pattern. This correlation is visible both in transcriptomes and proteomes. The same pattern is detectable by two independent methods, that is via emergence of founder genes by genomic phylostratigraphy, as well as via looking at recent evolutionary history by estimating evolutionary divergence rates of coding sequences.
2. *B. subtilis* biofilm genes important for multicellularity and with functions analogous to those important in animal development, have the highest expression during mid-period biofilms when the biofilms exhibit their typical wrinkled morphology.
3. *B. subtilis* biofilm growth at the functional level has clearly distinct hierarchical and stage-organized composition comparable to developmental processes in multicellular eukaryotes.

6. REFERENCES

Abzhanov, A. (2013) 'Von Baer's law for the ages: Lost and found principles of developmental evolution', *Trends in Genetics*, 29(12), pp. 712–722. doi:10.1016/j.tig.2013.09.004.

Aguilar, C. *et al.* (2007) 'Thinking about *Bacillus subtilis* as a multicellular organism', *Current Opinion in Microbiology*, 10(6), pp. 638–643. doi:10.1016/j.mib.2007.09.006.

Aguilar, C. *et al.* (2010) 'KinD Is a Checkpoint Protein Linking Spore Formation to Extracellular-Matrix Production in *Bacillus subtilis* Biofilms', *mBio*. Edited by E.P. Greenberg, 1(1). doi:10.1128/mBio.00035-10.

Altschul, S.F. *et al.* (1990) 'Basic local alignment search tool', *Journal of Molecular Biology*, 215(3), pp. 403–410. doi:10.1016/S0022-2836(05)80360-2.

Andrews, S. (2010) FastQC: A Quality Control Tool for High Throughput Sequence Data [Online]. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

Asally, M. *et al.* (2012) 'Localized cell death focuses mechanical forces during 3D patterning in a biofilm', *Proceedings of the National Academy of Sciences*, 109(46), pp. 18891–18896. doi:10.1073/pnas.1212429109.

Baer, K.E. von (1828) 'Über Entwickelungsgeschichte der Thiere: Beobachtung und Reflexion', *Bornträger* [Preprint].

Bogdanović, O. *et al.* (2016) 'Active DNA demethylation at enhancers during the vertebrate phylotypic period', *Nature Genetics*, 48(4), pp. 417–426. doi:10.1038/ng.3522.

Bushnell, B. (2014) BBMap: A Fast, Accurate, Splice-Aware Aligner. Available at: https://sourceforge.net/projects/bbmap/.

Cao, J. *et al.* (2019) 'The single-cell transcriptional landscape of mammalian organogenesis', *Nature*, 566(7745), pp. 496–502. doi:10.1038/s41586-019-0969-x.

Chai, Y. *et al.* (2010) 'An epigenetic switch governing daughter cell separation in *Bacillus subtilis*', *Genes & Development*, 24(8), pp. 754–765. doi:10.1101/gad.1915010.

Chai, Y., Kolter, R. and Losick, R. (2009) ‘Paralogous antirepressors acting on the master regulator for biofilm formation in *Bacillus subtilis*: Paralogous antirepressors acting on the master regulator’, *Molecular Microbiology*, 74(4), pp. 876–887. doi:10.1111/j.1365-2958.2009.06900.x.

Cheng, X. *et al.* (2015) ‘A “developmental hourglass” in fungi’, *Molecular Biology and Evolution*, 32(6), pp. 1556–1566. doi:10.1093/molbev/msv047.

Chu, F. *et al.* (2008) ‘A novel regulatory protein governing biofilm formation in *Bacillus subtilis*’, *Molecular Microbiology*, 68(5), pp. 1117–1127. doi:10.1111/j.1365-2958.2008.06201.x.

Comeron, J.M. (1995) ‘A method for estimating the numbers of synonymous and nonsynonymous substitutions per site’, *Journal of Molecular Evolution*, 41(6), pp. 1152–1159. doi:10.1007/BF00173196.

Comte, A., Roux, J. and Robinson-Rechavi, M. (2010) ‘Molecular signaling in zebrafish development and the vertebrate phylotypic period’, *Evolution and Development*, 12(2), pp. 144–156. doi:10.1111/j.1525-142X.2010.00400.x.

Cox, J. and Mann, M. (2008) ‘MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification’, *Nature Biotechnology*, 26(12), pp. 1367–1372. doi:10.1038/nbt.1511.

Domazet-Lošo, T., Brajković, J. and Tautz, D. (2007) ‘A phylostratigraphy approach to uncover the genomic history of major adaptations in metazoan lineages’, *Trends in Genetics*, 23(11), pp. 533–539. doi:10.1016/J.TIG.2007.08.014.

Domazet-Lošo, T. and Tautz, D. (2010) ‘A phylogenetically based transcriptome age index mirrors ontogenetic divergence patterns’, *Nature*, 468(7325), pp. 815–819. doi:10.1038/nature09632.

Drost, H.G. *et al.* (2015) ‘Evidence for active maintenance of phylotranscriptomic hourglass patterns in animal and plant embryogenesis’, *Molecular Biology and Evolution*, 32(5), pp. 1221–1231. doi:10.1093/molbev/msv012.

Drost, H.-G. *et al.* (2016) ‘Post-embryonic Hourglass Patterns Mark Ontogenetic Transitions in Plant Development’, *Molecular Biology and Evolution*, 33(5), pp. 1158–1163. doi:10.1093/molbev/msw039.

Drost, H.-G. *et al.* (2017) ‘Cross-kingdom comparison of the developmental hourglass’, *Current Opinion in Genetics & Development*, 45, pp. 69–75. doi:10.1016/j.gde.2017.03.003.

Drost, H.-G. *et al.* (2018) ‘myTAI: evolutionary transcriptomics with R’, *Bioinformatics*. Edited by Z. Bar-Joseph, 34(9), pp. 1589–1590. doi:10.1093/bioinformatics/btx835.

Duboule, D. (1994) ‘Temporal colinearity and the phylotypic progression: A basis for the stability of a vertebrate Bauplan and the evolution of morphologies through heterochrony’, *Development*, 120(SUPPL.), pp. 135–142.

Elek, A., Kuzman, M. and Vlahoviček, K. (2020) coRdon: Codon Usage Analysis and Prediction of Gene Expressivity. Available at: https://github.com/BioinfoHR/coRdon.

Futo, M. *et al.* (2021) ‘Embryo-Like Features in Developing *Bacillus subtilis* Biofilms’, *Molecular Biology and Evolution*. Edited by N. Perna, 38(1), pp. 31–47. doi:10.1093/molbev/msaa217.

van Gestel, J., Vlamakis, H. and Kolter, R. (2015) ‘Division of Labor in Biofilms: the Ecology of Cell Differentiation’, *Microbiology Spectrum*, 3(2), pp. 1–24. doi:10.1128/microbiolspec.mb-0002-2014.

Haeckel, E. (1868) ‘Natürliche Schöpfungsgeschichte’, *Archiv für Anthropologie* [Preprint].

Hall-Stoodley, L., Costerton, J.W. and Stoodley, P. (2004) ‘Bacterial biofilms: From the natural environment to infectious diseases’, *Nature Reviews Microbiology*, 2(2), pp. 95–108. doi:10.1038/nrmicro821.

Hashuel, R. and Ben-Yehuda, S. (2019) ‘Aging of a Bacterial Colony Enforces the Evolvement of Nondifferentiating Mutants’, *mBio*. Edited by C.S. Harwood, 10(5). doi:10.1128/mBio.01414-19.

Irie, N. and Kuratani, S. (2011) ‘Comparative transcriptome analysis reveals vertebrate phylotypic period during organogenesis’, *Nature Communications*, 2(1). doi:10.1038/ncomms1248.

Kalamara, M. *et al.* (2018) 'Social behaviours by *Bacillus subtilis*: quorum sensing, kin discrimination and beyond', *Molecular Microbiology*, 110(6), pp. 863–878. doi:10.1111/mmi.14127.

Kalinka, A.T. *et al.* (2010) 'Gene expression divergence recapitulates the developmental hourglass model', *Nature*, 468(7325), pp. 811–816. doi:10.1038/nature09634.

Kalinka, A.T. and Tomancak, P. (2012) 'The evolution of early animal embryos: Conservation or divergence?', *Trends in Ecology and Evolution*, 27(7), pp. 385–393. doi:10.1016/j.tree.2012.03.007.

Kearns, D.B. *et al.* (2004) 'A master regulator for biofilm formation by *Bacillus subtilis*: Master regulator for biofilm formation', *Molecular Microbiology*, 55(3), pp. 739–749. doi:10.1111/j.1365-2958.2004.04440.x.

Khalturin, K. *et al.* (2009) 'More than just orphans: are taxonomically-restricted genes important in evolution?', *Trends in Genetics*, 25(9), pp. 404–413. doi:10.1016/j.tig.2009.07.006.

Kobayashi, K. (2007) 'Gradual activation of the response regulator DegU controls serial expression of genes for flagellum formation and biofilm formation in *Bacillus subtilis*', *Molecular Microbiology*, 66(2), pp. 395–409. doi:10.1111/j.1365-2958.2007.05923.x.

Kolodkin-Gal, I. *et al.* (2010) 'D -Amino Acids Trigger Biofilm Disassembly', *Science*, 328(5978), pp. 627–629. doi:10.1126/science.1188628.

Lawrence, M. *et al.* (2013) 'Software for Computing and Annotating Genomic Ranges', *PLoS Computational Biology*. Edited by A. Prlic, 9(8), p. e1003118. doi:10.1371/journal.pcbi.1003118.

Levin, M. *et al.* (2012) 'Developmental Milestones Punctuate Gene Expression in the *Caenorhabditis* Embryo', *Developmental Cell*, 22(5), pp. 1101–1108. doi:10.1016/j.devcel.2012.04.004.

Levin, M. *et al.* (2016) 'The mid-developmental transition and the evolution of animal body plans', *Nature*, 531(7596), pp. 637–641. doi:10.1038/nature16994.

Li, B. *et al.* (2009) 'RNA-Seq gene expression estimation with read mapping uncertainty', *Bioinformatics*, 26(4), pp. 493–500. doi:10.1093/bioinformatics/btp692.

Li, H. *et al.* (2009) 'The Sequence Alignment/Map format and SAMtools', *Bioinformatics*, 25(16), pp. 2078–2079. doi:10.1093/bioinformatics/btp352.

Liu, J. *et al.* (2015) 'Metabolic co-dependence gives rise to collective oscillations within biofilms', *Nature*, 523(7562), pp. 550–554. doi:10.1038/nature14660.

Liu, Y., Beyer, A. and Aebersold, R. (2016) 'On the Dependency of Cellular Protein Levels on mRNA Abundance', *Cell*, 165(3), pp. 535–550. doi:10.1016/j.cell.2016.03.014.

Logan, B.E. (2009) 'Exoelectrogenic bacteria that power microbial fuel cells', *Nature Reviews Microbiology*, 7(5), pp. 375–381. doi:10.1038/nrmicro2113.

Lopez, D. *et al.* (2009) 'Structurally diverse natural products that cause potassium leakage trigger multicellularity in *Bacillus subtilis*', *Proceedings of the National Academy of Sciences*, 106(1), pp. 280–285. doi:10.1073/pnas.0810940106.

López, D. and Kolter, R. (2010) 'Extracellular signals that define distinct and coexisting cell fates in *Bacillus subtilis*', *FEMS Microbiology Reviews*, 34(2), pp. 134–149. doi:10.1111/j.1574-6976.2009.00199.x.

López, D., Vlamakis, H. and Kolter, R. (2010) 'Biofilms.', *Cold Spring Harbor perspectives in biology*, 2(7). doi:10.1101/cshperspect.a000398.

Love, M.I., Huber, W. and Anders, S. (2014) 'Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2', *Genome Biology*, 15(12), p. 550. doi:10.1186/s13059-014-0550-8.

Marlow, V.L. *et al.* (2014) 'The prevalence and origin of exoprotease-producing cells in the *Bacillus subtilis* biofilm', *Microbiology*, 160(1), pp. 56–66. doi:10.1099/mic.0.072389-0.

McDowell, I.C. *et al.* (2018) 'Clustering gene expression time series data using an infinite Gaussian process mixture model', *PLOS Computational Biology*. Edited by Q. Nie, 14(1), p. e1005896. doi:10.1371/journal.pcbi.1005896.

de Mendoza, A. *et al.* (2013) 'Transcription factor evolution in eukaryotes and the assembly of the regulatory toolkit in multicellular lineages', *Proceedings of the National Academy of Sciences*, 110(50), pp. E4858–E4866. doi:10.1073/pnas.1311818110.

Monds, R.D. and O'Toole, G.A. (2009) 'The developmental model of microbial biofilms: ten years of a paradigm up for review', *Trends in Microbiology*, 17(2), pp. 73–87. doi:10.1016/j.tim.2008.11.001.

Morgan, M. *et al.* (2017) Rsamtools: Binary alignment (BAM), FASTA, variant call (BCF), and tabix file import. Available from: http://bioconductor.org/packages/release/bioc/html/Rsamtools.html.

Niklas, K.J. (2014) 'The evolutionary-developmental origins of multicellularity', *American Journal of Botany*, 101(1), pp. 6–25. doi:10.3732/ajb.1300314.

Nye, T.M. *et al.* (2020) 'Methyltransferase DnmA is responsible for genome-wide N6-methyladenosine modifications at non-palindromic recognition sites in *Bacillus subtilis*', *Nucleic Acids Research*, 48(10), pp. 5332–5348. doi:10.1093/nar/gkaa266.

Peshkin, L. *et al.* (2015) 'On the Relationship of Protein and mRNA Dynamics in Vertebrate Embryonic Development', *Developmental Cell*, 35(3), pp. 383–394. doi:10.1016/j.devcel.2015.10.010.

Prindle, A. *et al.* (2015) 'Ion channels enable electrical communication in bacterial communities', *Nature*, 527(7576), pp. 59–63. doi:10.1038/nature15709.

Quint, M. *et al.* (2012) 'A transcriptomic hourglass in plant embryogenesis', *Nature*, 490(7418), pp. 98–101. doi:10.1038/nature11394.

R Development Core Team (2008) R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. Available from: http://www.R-project.org.

Raj, B. *et al.* (2020) 'Emergence of Neuronal Diversity during Vertebrate Brain Development', *Neuron*, 108(6), pp. 1058-1074.e6. doi:10.1016/j.neuron.2020.09.023.

Ratnayake-Lecamwasam, M. *et al.* (2001) '*Bacillus subtilis* CodY represses early-stationary-phase genes by sensing GTP levels', *Genes & Development*, 15(9), pp. 1093–1103. doi:10.1101/gad.874201.

Richardson, M.K. (1999) 'Vertebrate evolution: The developmental origins of adult variation', *BioEssays*, 21(7), pp. 604–613. doi:10.1002/(SICI)1521-1878(199907)21:7<604::AID-BIES9>3.0.CO;2-U.

Røder, H.L., Sørensen, S.J. and Burmølle, M. (2016) 'Studying Bacterial Multispecies Biofilms: Where to Start?', *Trends in Microbiology*, 24(6), pp. 503–513. doi:10.1016/j.tim.2016.02.019.

Romaní, A.M. *et al.* (2008) 'Relevance of polymeric matrix enzymes during biofilm formation', *Microbial Ecology*, 56(3), pp. 427–436. doi:10.1007/s00248-007-9361-8.

Roux, J. and Robinson-Rechavi, M. (2008) 'Developmental constraints on vertebrate genome evolution', *PLoS Genetics*, 4(12). doi:10.1371/journal.pgen.1000311.

Shapiro, J.A. (1998) 'Thinking About Bacterial Populations As Multicellular Organisms', *Annual Review of Microbiology*, 52(1), pp. 81–104. doi:10.1146/annurev.micro.52.1.81.

Sharp, P.M. (2005) 'Variation in the strength of selected codon usage bias among bacteria', *Nucleic Acids Research*, 33(4), pp. 1141–1153. doi:10.1093/nar/gki242.

Shi, L. *et al.* (2020) 'Evolutionary Analysis of the *Bacillus subtilis* Genome Reveals New Genes Involved in Sporulation', *Molecular Biology and Evolution*. Edited by D. Agashe, 37(6), pp. 1667–1678. doi:10.1093/molbev/msaa035.

Sierro, N. *et al.* (2008) 'DBTBS: A database of transcriptional regulation in *Bacillus subtilis* containing upstream intergenic conservation information', *Nucleic Acids Research*, 36(SUPPL. 1), pp. 93–96. doi:10.1093/nar/gkm910.

Singh, R., Paul, D. and Jain, R.K. (2006) 'Biofilms: implications in bioremediation', *Trends in Microbiology*, 14(9), pp. 389–397. doi:10.1016/j.tim.2006.07.001.

Slack, J.M.W., Holland, P.W.H. and Graham, C.F. (1993) 'The zootype and the phylotypic stage', *Nature*, 361(6412), pp. 490–492. doi:10.1038/361490a0.

Spacapan, M., Danevčič, T. and Mandic-Mulec, I. (2018) 'ComX-Induced Exoproteases Degrade ComX in *Bacillus subtilis* PS-216', *Frontiers in Microbiology*, 9, p. 105. doi:10.3389/fmicb.2018.00105.

Srinivasan, S. *et al.* (2018) 'Matrix Production and Sporulation in *Bacillus subtilis* Biofilms Localize to Propagating Wave Fronts', *Biophysical Journal*, 114(6), pp. 1490–1498. doi:10.1016/j.bpj.2018.02.002.

Stoodley, P. *et al.* (2002) 'Biofilms as Complex Differentiated Communities', *Annual Review of Microbiology*, 56(1), pp. 187–209. doi:10.1146/annurev.micro.56.012302.160705.

Supek, F. and Vlahoviček, K. (2005) 'Comparison of codon usage measures and their applicability in prediction of microbial gene expressivity', *BMC Bioinformatics*, 6(iii), pp. 1–15. doi:10.1186/1471-2105-6-182.

Suyama, M., Torrents, D. and Bork, P. (2006) 'PAL2NAL: Robust conversion of protein sequence alignments into the corresponding codon alignments', *Nucleic Acids Research*, 34(WEB. SERV. ISS.), pp. 609–612. doi:10.1093/nar/gkl315.

Tautz, D. and Domazet-Lošo, T. (2011) 'The evolutionary origin of orphan genes', *Nature Reviews Genetics*, 12(10), pp. 692–702. doi:10.1038/nrg3053.

Tomancak, P. *et al.* (2007) 'Global analysis of patterns of gene expression during *Drosophila* embryogenesis', *Genome Biology*, 8(7), p. R145. doi:10.1186/gb-2007-8-7-r145.

Tyanova, S., Temu, T. and Cox, J. (2016) 'The MaxQuant computational platform for mass spectrometry-based shotgun proteomics', *Nature Protocols*, 11(12), pp. 2301–2319. doi:10.1038/nprot.2016.136.

Vlamakis, H. *et al.* (2008) 'Control of cell fate by the formation of an architecturally complex bacterial community', *Genes & Development*, 22(7), pp. 945–953. doi:10.1101/gad.1645008.

Vlamakis, H. *et al.* (2013) 'Sticking together: Building a biofilm the *Bacillus subtilis* way', *Nature Reviews Microbiology*, 11(3), pp. 157–168. doi:10.1038/nrmicro2960.

Wickham, H. (2016) ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York Available from: http://ggplot2.org.

Wolpert, L. *et al.* (2007) *Principles of Development*. Third. Oxford University Press.

Xu, J. and Zhang, S. (2015) 'Mitogen-activated protein kinase cascades in signaling plant growth and development', *Trends in Plant Science*, 20(1), pp. 56–64. doi:10.1016/j.tplants.2014.10.001.

Yanai, I. (2018) 'Development and Evolution through the Lens of Global Gene Regulation', *Trends in Genetics*, 34(1), pp. 11–20. doi:10.1016/j.tig.2017.09.011.

Yekutieli, D. and Benjamini, Y. (1999) 'Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics', *Journal of Statistical Planning and Inference*, 82(1–2), pp. 171–196. doi:10.1016/s0378-3758(99)00041-5.

Zeisel, A. *et al.* (2018) 'Molecular Architecture of the Mouse Nervous System', *Cell*, 174(4), pp. 999-1014.e22. doi:10.1016/j.cell.2018.06.021.

Zhang, W. and Liu, H.T. (2002) 'MAPK signal pathways in the regulation of cell proliferation in mammalian cells', *Cell Research*, 12(1), pp. 9–18. doi:10.1038/sj.cr.7290105.

Zhu, B. and Stülke, J. (2018) 'SubtiWiki in 2018: From genes and proteins to functional network annotation of the model organism *Bacillus subtilis'*, *Nucleic Acids Research*, 46(D1), pp. D743–D748. doi:10.1093/nar/gkx908.

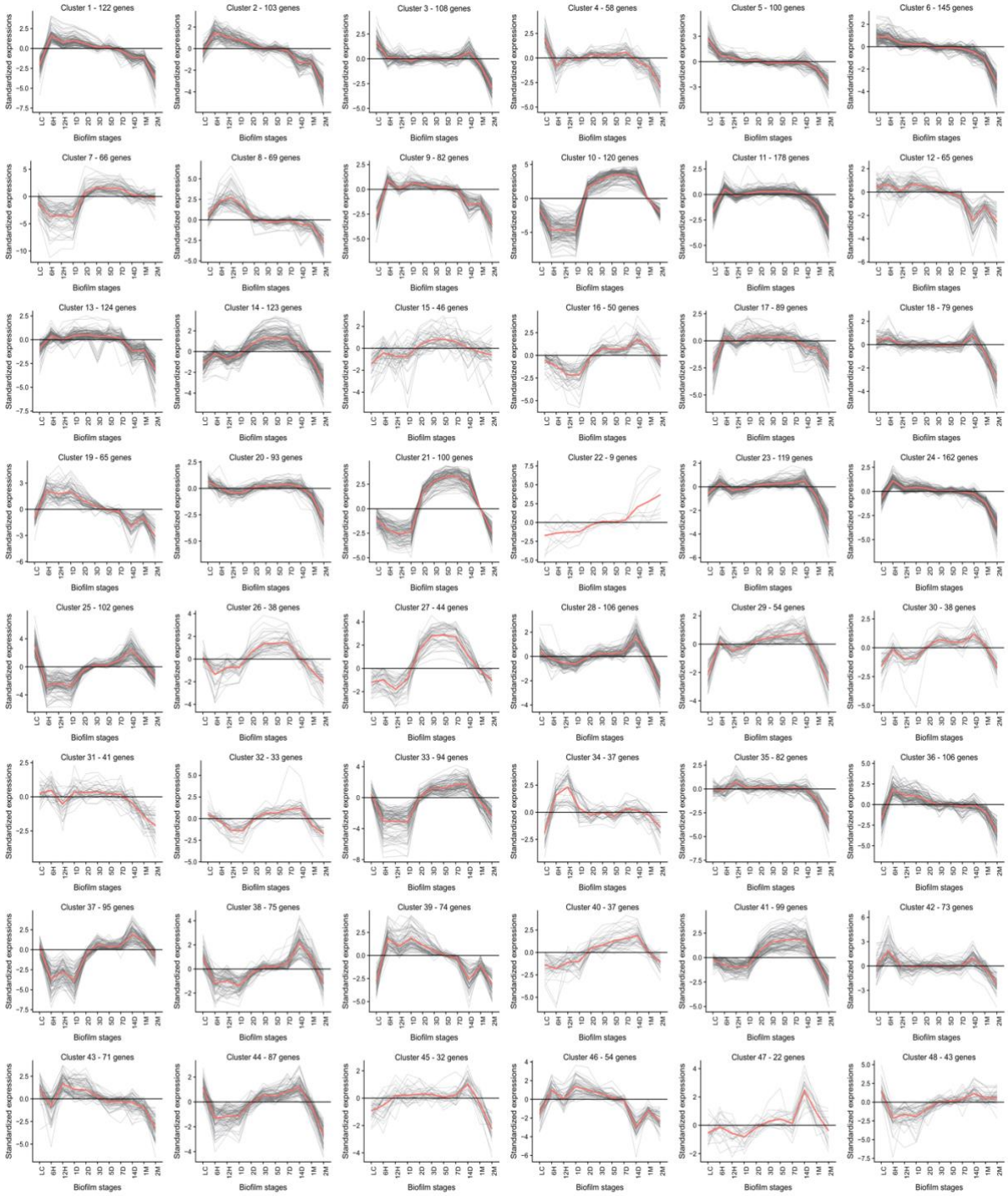
7. APPENDICES

Appendix 1. Details of mapping *B. subtilis* subsp. *subtilis* str. NCIB 3610 sequences onto the reference genome shows low variation between the samples. Percentage of mapped reads is calculated as a ratio between the total number of mapped reads and number of reads used for mapping. Coverage is calculated as a ratio between the total number of mapped reads and number of protein coding nucleotides in the *B. subtilis* genome. Table obtained from Futo *et al.* (2021).

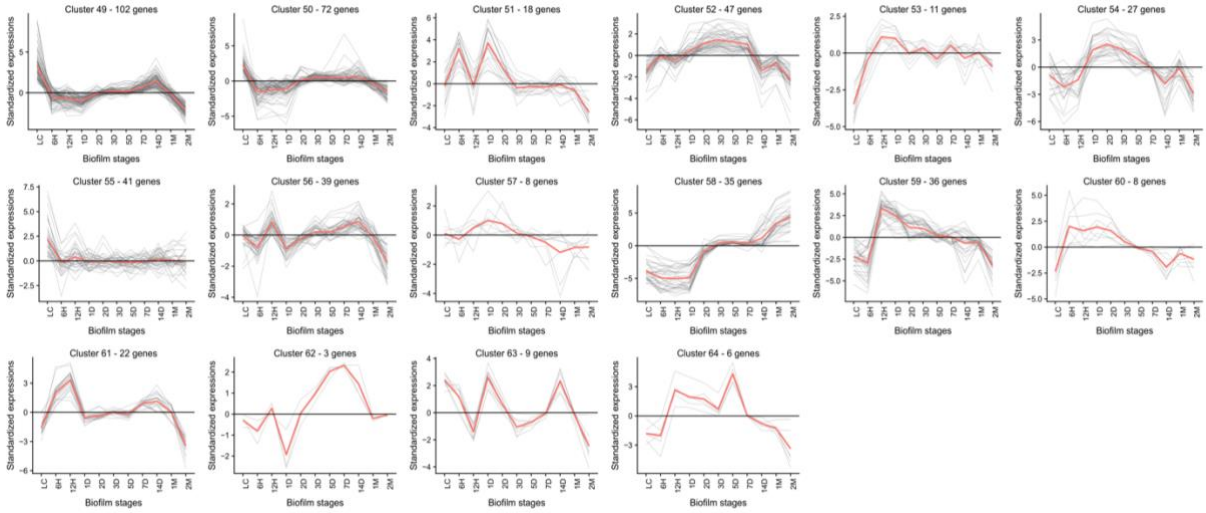
Timepoint_replicate	Reads used for mapping	Total number of mapped reads	Percentage of mapped reads (%)	Number of bases used for mapping	Total number of mapped bases	Protein coding nucleotides in the *B. subtilis* genome	Coverage, x
6H_rep1	52927176	51012866	96.38312462	3969538200	3681683550	3770118	976.5433204
6H_rep2	58255952	55260848	94.8587159	4369196400	3914336550	3770118	1038.25306
6H_rep3	48686198	45677793	93.82082577	3651464850	3181296900	3770118	843.8189203
12H_rep1	46708936	44994189	96.32886735	3503170200	3251213850	3770118	862.3639499
12H_rep2	48602422	46361257	95.38877918	3645181650	3324079950	3770118	881.6912229
12H_rep3	48161416	45965801	95.44113279	3612106200	3287378100	3770118	871.9562889
1D_rep1	61419704	58224065	94.79704591	4606477800	4129674000	3770118	1095.369959
1D_rep2	42125932	38878345	92.29076522	3159444900	2718387450	3770118	721.0351108
1D_rep3	51561796	49812612	96.60759683	3867134700	3638044050	3770118	964.9682185
2D_rep1	68258838	65382486	95.78611051	5119412850	4765350000	3770118	1263.979005
2D_rep2	63543056	61230907	96.36128769	4765729200	4441112550	3770118	1177.977069
2D_rep3	63667114	60957107	95.74347441	4775033550	4403622150	3770118	1168.032977
3D_rep1	62722252	60791425	96.9216236	5017780160	4662173920	3770118	1236.612202
3D_rep2	56239654	53240177	94.66661548	4499172320	3958504800	3770118	1049.968409
3D_rep3	44836234	42728071	95.2980819	3586898720	3251291680	3770118	862.3845938
5D_rep1	45391352	43945342	96.81434913	3404351400	3202447950	3770118	849.4291027
5D_rep2	43344730	39893734	92.03825702	3250854750	2720504400	3770118	721.5966185
5D_rep3	63143302	58983408	93.41197899	4735747650	4108402950	3770118	1089.727948
7D_rep1	50079410	48303497	96.45380607	4006352800	3659820320	3770118	970.7442367
7D_rep2	42337894	39946285	94.35113849	3387031520	3039407520	3770118	806.1836579
7D_rep3	44780284	41859068	93.47655767	3582422720	3150831040	3770118	835.7380432

Appendix 1. – continued.

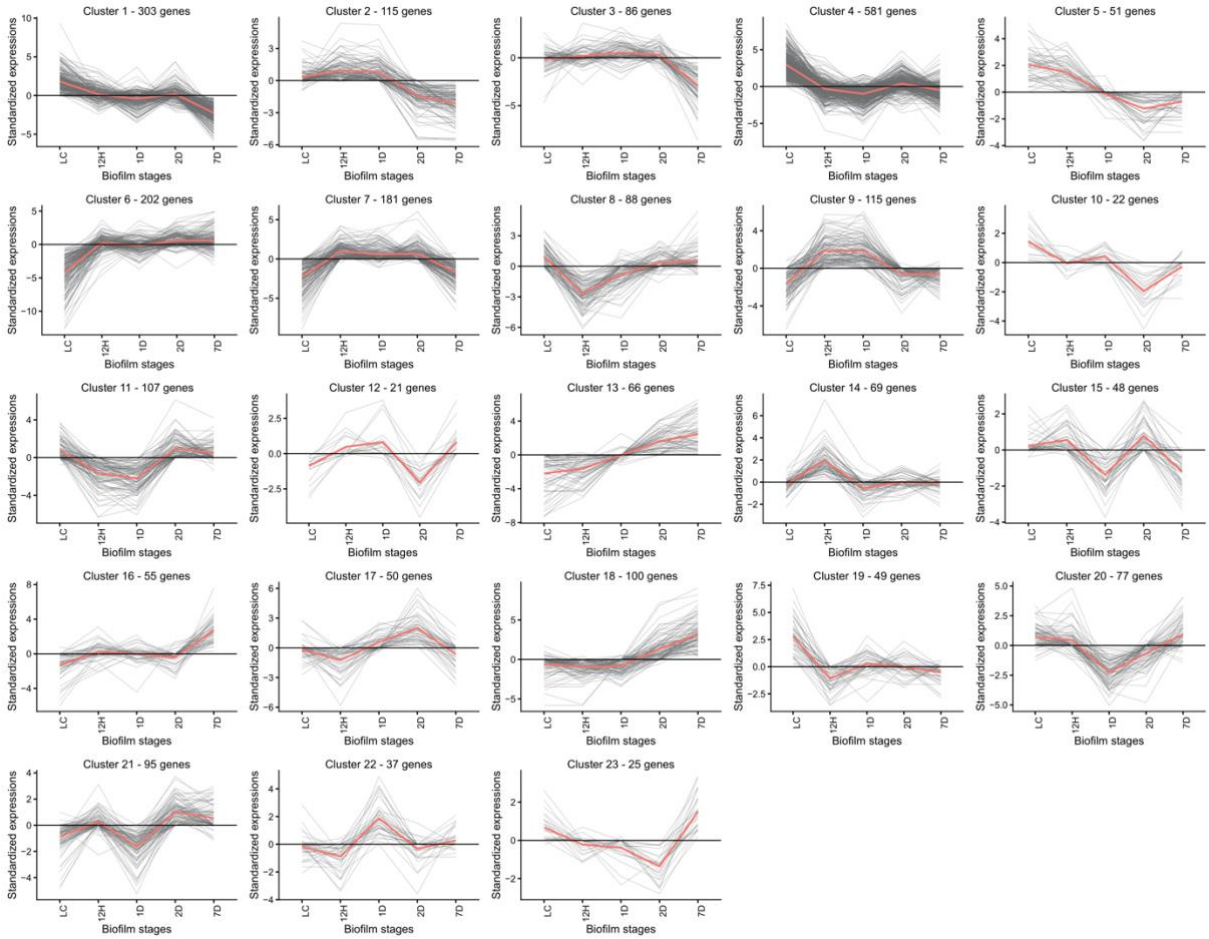
14D_rep1	48423984	46551430	96.13300302	3631798800	3347284800	3770118	887.8461629
14D_rep2	41925954	32727454	78.06012953	3354076320	2463176800	3770118	653.342097
14D_rep3	38767430	33617264	86.71522461	3101394400	2552777280	3770118	677.1080587
1M_rep1	49555266	45283206	91.37920075	3716644950	3152953500	3770118	836.3010123
1M_rep2	51696532	47261102	91.4202562	3877239900	3343707900	3770118	886.8974128
1M_rep3	51249678	47967147	93.59502122	3843725850	3511213200	3770118	931.3271362
2M_rep1	29477002	22767316	77.23755625	2358160160	1641277120	3770118	435.3383952
LC_rep1	42397708	40565398	95.67828053	3391816640	3106647840	3770118	824.0187283
LC_rep2	34339860	31882798	92.84486891	2747188800	2395698400	3770118	635.4438774
LC_rep3	48191656	46720660	96.94761267	3855332480	3662362080	3770118	971.4184224
AVG	**49768345.87**	**46735259.94**	**93.45971898**	**3819092930**	**3408602019**		**904.1101683**
SUM	**1542818722**	**1448793058**					



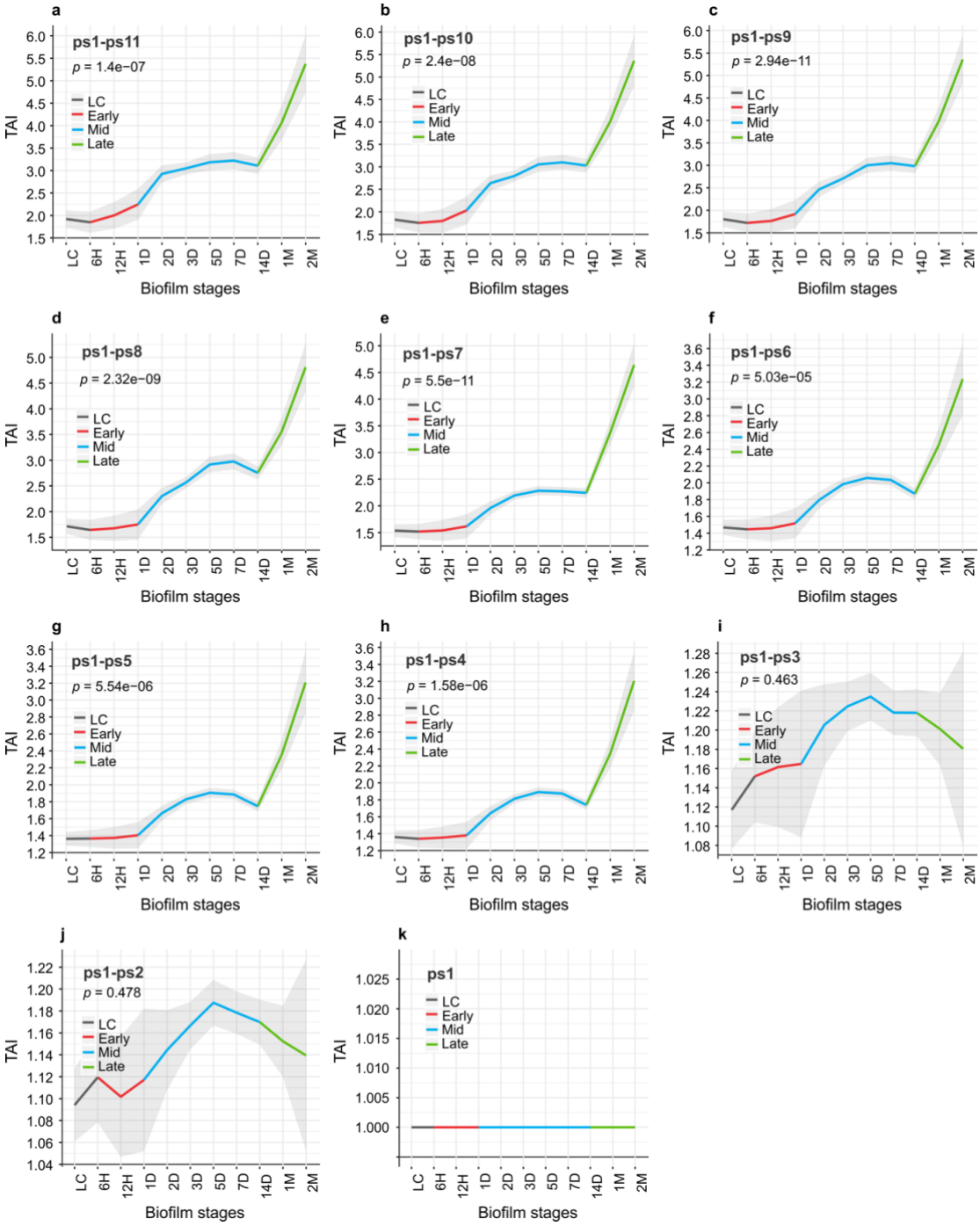
Appendix 2. *Bacillus subtilis* subsp. *subtilis* str. NCIB 3610 biofilm transcriptome clusters show that biofilm growth is not a continuous process. Gray lines represent transcriptome standardized expression values of individual genes, red line represents the average transcriptome standardized expression value of genes in a cluster, and black horizontal line represents the median of standardized transcriptome expression values. Transcriptome clusters are obtained by DP_GP_cluster with the maximum Gibbs sampling iterations set to 500. The list of genes in individual clusters can be seen in Futo *et al.* (2021).



Appendix 2. – continued.



Appendix 3. *Bacillus subtilis* subsp. *subtilis* str. NCIB 3610 proteome clusters show that biofilm growth is not a continuous process. Gray lines represent proteome standardized expression values of individual proteins, red line represents the average proteome standardized expression value of proteins in a cluster, and black horizontal line represents the median of standardized proteome expression values. Proteome clusters are obtained by DP_GP_cluster with the maximum Gibbs sampling iterations set to 500. The list of genes in individual clusters can be seen in Futo *et al.* (2021).



Appendix 4. Recapitulation pattern is significant from the origin of Firmicutes at ps4. Transcriptome age index (TAI) was calculated on the reduced datasets by progressively removing genes from younger phylostrata. **a)** ps1-ps11 (n = 4,293); **b)** ps1-ps10 (n = 4,218); **c)** ps1-ps9 (n = 4,186); **d)** ps1-ps8 (n = 4,097); **e)** ps1-ps7 (n = 3,923); **f)** ps1-ps6 (n = 3,778); **g)** ps1-ps5 (n = 3,707); **h)** ps1-ps4 (n = 3,683); **i)** ps1-ps3 (n = 3,239); **j)** ps1-ps2 (n = 3,159) and **k)** ps1 (n = 2,562). Depicted *p* values are obtained by the flat line test and grey shaded areas represent ± one standard deviation estimated by permutation analysis (see Material and Methodology**)**. Early (red), mid (blue) and late (green) periods of biofilm growth are colour-coded. Figure obtained from Futo *et al.* (2021).

Appendix 5. Distribution of *Bacillus subtilis* subsp. s*ubtilis* NCIB 3610 genes on the phylostratigraphy maps made by BLASTp different e-value cut-offs. High e-value cut-offs push gene ages towards the older phylostrata and inflate the false positive rates, while high e-value cut-offs push gene ages in older phylostrata and inflate false negative rates. Regardless of the shifts in gene ages, the TAI pattern remains stable (see Figure 6).

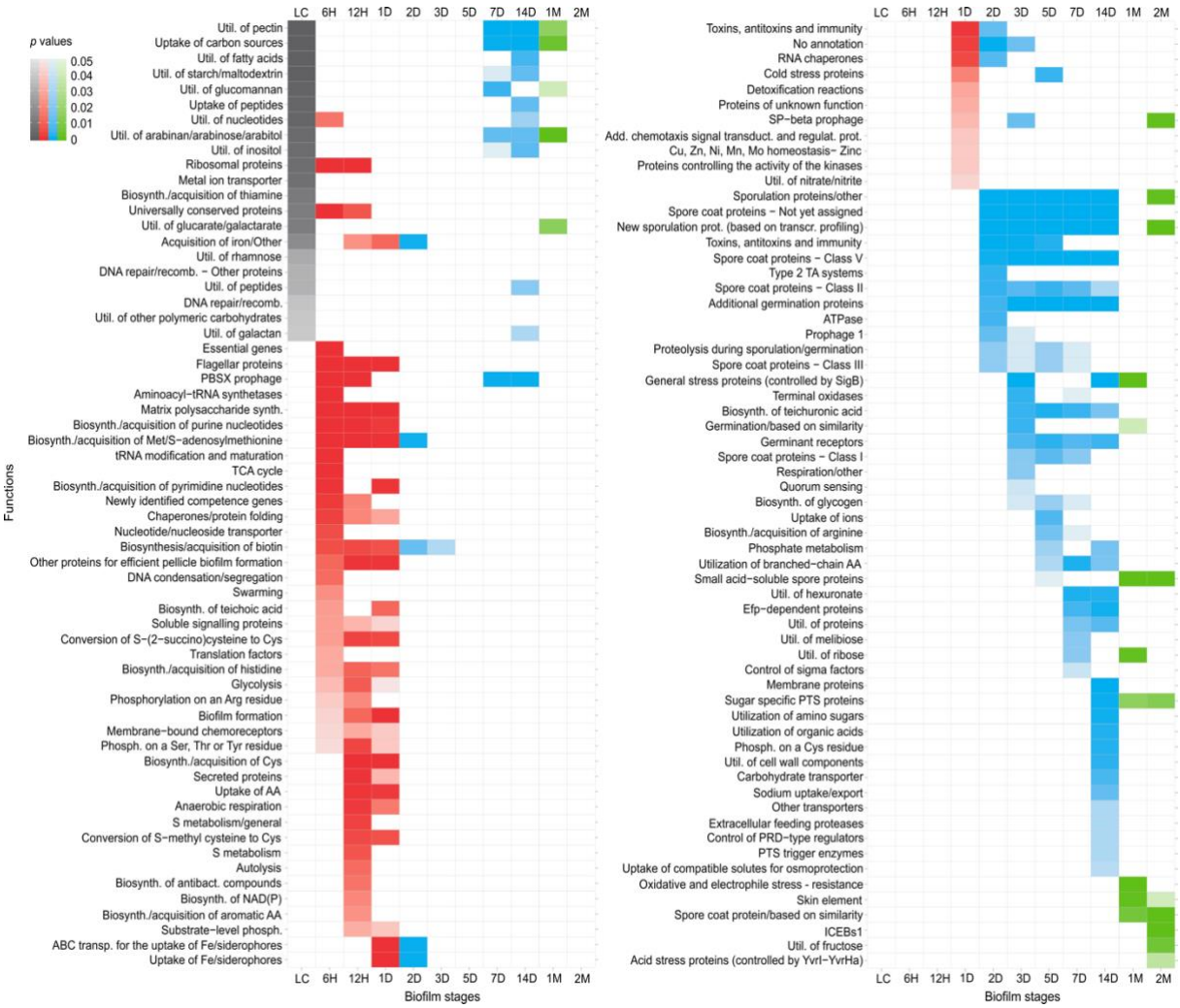
PS	1E-30	1E-20	1E-15	1E-10	1E-5	1E-2	1E-1	1	10
1	1426	1721	1927	2156	2444	2677	2818	3129	3940
2	798	823	793	727	637	555	496	471	215
3	122	111	94	90	88	73	82	114	38
4	635	561	538	515	456	413	382	269	47
5	28	32	35	34	26	23	24	15	6
6	89	89	79	68	75	69	59	43	5
7	197	181	176	177	155	139	130	77	22
8	520	407	331	257	197	164	143	77	16
9	83	98	111	109	97	83	75	55	16
10	49	60	49	36	31	29	28	20	3
11	164	142	117	102	83	72	64	38	11
12	60	50	42	35	27	21	18	11	4
total	4171	4275	4292	4306	4316	4318	4319	4319	4323

Appendix 6. Gene identities and transcriptome standardized expression values for categories "Transcription regulators", "Cell-cell signalling", "Protein kinases" and "Protein phosphatases" shown in Figure 4a-e. Table modified from Futo *et al.* (2021).

Gene symbol	LC	6H	12H	1D	2D	3D	5D	7D	14D	1M	2M	Category
SigA	1.099816	-0.25077	-0.56001	-0.63624	0	0.130907	0.214582	0.304288	0.150577	-0.16812	-2.22825	trans reg - sigma f
SigB	0.19176	-4.56653	-2.24133	-4.15026	-0.90346	0.625903	0	0.422927	1.953452	0.889038	-0.67328	trans reg - sigma f
SigD	-0.8658	0.452489	0.105341	-0.59638	-0.02321	-0.10113	0.127656	-0.08713	0.805358	1.177105	0	trans reg - sigma f
SigE	-0.71282	-2.50492	-3.0669	-2.14846	1.516482	3.140957	2.60214	3.630061	1.257064	0	-2.00727	trans reg - sigma f
SigF	1.048998	-2.27883	-0.70163	-1.38849	0.56312	1.48365	1.400539	2.373713	0	-1.34765	-3.10245	trans reg - sigma f
SigG	-1.70185	-1.72542	-2.49899	-2.60971	1.716411	3.127771	3.51438	3.479138	1.714212	0	-0.34743	trans reg - sigma f
SigH	-2.38716	-0.16965	-0.40517	0.019997	0.352636	0	0.207456	0.051885	0.147714	-0.1811	-1.75006	trans reg - sigma f
SigKN	-0.82363	-0.71038	-0.92603	-0.51913	1.4097	2.303667	2.588531	3.339532	2.614244	0	-2.39692	trans reg - sigma f
SigKC	-1.00551	0	-1.0604	-0.71997	1.473691	2.548477	3.005355	3.381454	2.672196	-0.17824	-2.12265	trans reg - sigma f
SigW	1.138363	-0.37657	-0.64429	0	0.635064	0.714912	0.745061	0.881942	-0.89784	-0.40399	-3.49136	trans reg - sigma f
SigX	-3.06856	0.853443	1.65836	0.802997	0.532591	0.162625	0	-0.15608	-2.65304	-1.39027	-3.93667	trans reg - sigma f
DegU	0.608282	1.435629	0.023123	0	0.121563	-0.22152	-0.06551	0.466997	-0.00357	-0.50823	-2.59853	trans reg - other
GerE	-1.55674	-3.39798	-3.77787	-4.03367	2.430828	3.144297	4.099659	3.914196	2.433121	0	-1.58768	trans reg - other
CodY	0.658681	0.47068	-0.07995	0	0.289571	0.467083	0.041953	-0.5809	-1.85096	-1.76614	-3.57907	trans reg - other
LexA	-1.63529	0.926172	-0.02432	0.381066	0.236538	0.052228	0.223182	0	-0.8428	-0.86117	-2.82402	trans reg - other
Hpr	-2.59102	0.733018	0	-0.22706	0.402521	0.518528	0.566167	0.573293	-0.13168	-0.00959	-2.17688	trans reg - other
CcpA	-0.88777	1.827833	0.371997	0.571976	0.397675	0	0.062468	-0.15312	-0.15594	-1.13287	-3.02552	trans reg - other
TnrA	-1.85697	2.300378	-0.43641	3.645359	1.407416	0.360779	0	0.250033	-0.86463	-0.58514	-2.66599	trans reg - other
Fur	-0.9294	0.870149	-0.29804	0	0.588895	0.230216	0.835036	0.78507	-0.18015	-0.72058	-3.66804	trans reg - other
PhoP	0	-0.91511	-0.95377	-1.24356	0.143755	0.624502	0.616279	0.854827	-0.00412	0.353527	-1.14234	trans reg - other
ResD	0.353164	-0.20052	0.871681	0.629885	0.946784	0.930701	0	-0.15867	-0.30513	-0.99228	-3.61585	trans reg - other
AbrB	-0.79771	2.39473	1.143551	1.43148	0.778317	0.143438	0	-1.17597	-2.61121	-0.62603	-1.96629	trans reg - other
SpoVT	-0.72101	0.249874	-0.30786	-0.86329	0.773284	1.976854	2.322464	1.874956	0	-0.35089	-2.12991	trans reg - other
ResD	0.353164	-0.20052	0.871681	0.629885	0.946784	0.930701	0	-0.15867	-0.30513	-0.99228	-3.61585	trans reg - other
Spo0A	0	-0.2889	-0.39172	-0.47053	0.563415	0.774108	0.749329	1.399536	0.721344	-0.13683	-2.43944	trans reg - other
ComK	-0.62246	-1.19016	-1.20241	-2.08174	0.802386	2.398436	1.735909	1.352545	2.576214	0	-2.05493	trans reg - other
PurR	-0.02975	0.706968	-0.25082	0	0.032531	0.155791	0.307756	0.383924	-0.12476	-1.5507	-3.87916	trans reg - other
Rok	-1.58784	1.17936	0	0.422805	0.255035	0.192654	-0.04414	0.279217	-0.73776	-0.89606	-3.05718	trans reg - other
SpoIIID	-2.22841	-5.10228	-4.86827	-5.01063	2.200982	3.426701	3.932143	4.242509	2.420727	0	-1.79145	trans reg - other
comX	-1.11566	0	-1.18103	0.476385	0.48569	0.979991	0.225098	0.56684	-2.05329	-1.80079	-2.31366	cell cell sign - com
comQ	0	-0.46227	-0.85078	0.213655	0.821303	1.180291	0.984581	0.728228	-2.81783	-1.04878	-2.14494	cell cell sign - com
comP	-1.11501	0.113287	-0.20167	0.461646	0.35788	0.542855	0	0.053557	-0.40582	-0.94523	-2.26156	cell cell sign - com
comA	-1.4602	0	-0.40199	-0.13659	0.413865	0.398344	0.380261	0.627237	0.436459	-0.57055	-3.13934	cell cell sign - com
rapJ	-0.29453	0.65759	0.941885	0.978035	0.918646	0.365075	0	-0.13874	-0.98336	-0.76756	-2.24531	cell cell sign - rap
rapC	-1.21615	2.335645	-1.14712	-1.31909	1.043432	1.59807	1.218877	1.157051	0	-0.58944	-2.95531	cell cell sign - rap
rapI	-1.30126	0.998837	-0.6248	0.972828	0.600125	0.289668	0.015467	0	-2.75366	-1.79617	-2.46105	cell cell sign - rap
rapH	2.93347	-2.10697	-0.65958	-0.74977	0.253411	0.538349	0	0.395494	0.282748	-0.25986	-2.92229	cell cell sign - rap
rapA	4.93363	-3.07831	-2.89356	-4.34778	-1.8626	0	-0.19159	1.021603	2.452611	2.497477	0.197378	cell cell sign - rap
rapK	-3.47755	1.020281	0.042604	1.876368	1.348318	1.11024	-0.00968	0	-3.83322	-1.38625	-2.67977	cell cell sign - rap
rapE	-1.27919	0	-2.10852	-1.23861	-0.06993	1.008483	0.387274	1.166708	0.739897	0.561649	-0.49191	cell cell sign - rap
rapD	-1.01313	-0.1281	0.189461	-0.0343	0	0.390166	0.061987	0.31232	0.178103	-0.51169	-2.67088	cell cell sign - rap
rapB	-1.45468	0.022459	-0.21582	0.229854	0.338669	0.218853	0	0.01251	-0.87036	-0.68519	-2.91759	cell cell sign - rap
rapF	1.994238	2.676807	-1.48967	-1.08808	-0.06366	1.356047	0.390709	1.150483	0	-0.48624	-1.13493	cell cell sign - rap
rapG	1.333812	-0.73357	0.111828	-0.32171	-0.02438	0.186924	0.125283	0.48226	0	-1.02713	-3.57291	cell cell sign - rap
rapP	1.671786	0	0.589203	1.447546	1.1142	0.661725	-0.18633	-0.34296	-3.23934	-1.5459	-3.34561	cell cell sign - rap
phrC	-2.16993	0.259362	2.192039	2.727672	2.24826	0.848774	-0.19742	0	-0.9039	-0.78359	-4.5335	cell cell sign - phr
phrI	-0.34732	1.009348	0	1.106338	0.63568	0.534152	0.302833	-0.98196	-1.38505	-2.5988	-2.5988	cell cell sign - phr
phrH	2.488154	-2.46229	-0.62861	-1.0428	0.0901	0.328082	0	0.184534	1.022384	-0.40399	-2.00352	cell cell sign - phr
phrA	5.147402	-3.5697	-2.63767	-4.2457	-1.65966	0.525924	0	1.08662	2.518037	2.22107	-0.58061	cell cell sign - phr
phrK	-1.62959	0.506123	0	1.24507	1.297267	1.2172	-0.25657	0.247664	-0.94815	-0.99066	-2.94094	cell cell sign - phr
phrE	1.091813	-0.35226	-0.23866	0.310256	0.318312	0.596459	0	-0.03764	0.17216	-0.1676	-1.90286	cell cell sign - phr
phrF	0	1.190126	-0.98084	-0.13729	0.315126	0.452545	-0.35683	0.147428	0.641786	-1.22611	-2.94051	cell cell sign - phr
phrG	1.401679	-0.9543	0.510844	0.19104	0.363713	0	-0.20408	0.205058	-0.50493	-0.80047	-3.8368	cell cell sign - phr
walK	-0.31181	0.382686	0.994409	-0.1095	0.290554	-0.50299	0.035369	0	0.005475	-0.23927	-2.35159	prot kin
yxjM	2.176514	0.590926	0.221986	0.340742	0	-0.39915	-0.49323	-0.31657	0.381746	-1.09665	-4.01552	prot kin
ycbM	-1.20763	-0.00303	-0.93555	-0.54553	0.013039	1.064823	0.605239	0.46305	2.183966	0	-2.33323	prot kin
ydiB	0	0.967559	0.126768	-0.16555	0.313422	-0.17176	0.162685	0.030987	-0.06511	-0.64819	-2.77249	prot kin
mcsB	0	-2.59088	-1.27905	-2.08166	-0.4728	0.346724	0.34905	0.681791	1.041205	0.821458	-1.11266	prot kin
ywqC	-2.44833	1.439084	0.562087	0.90671	0.609771	-0.14205	0.218125	-0.10224	0	-1.17755	-3.90378	prot kin
yvrG	0.160277	0.972379	0.154385	0.458677	0.415467	0	-0.0908	-0.0132	-0.65219	-1.18471	-3.33357	prot kin
rsbW	0.254584	-4.36917	-2.49874	-4.33621	-0.87291	0.594619	0	0.040652	0.890928	0.373505	-1.98719	prot kin
spoIIAB	0.417091	-2.48865	-0.80994	-1.50162	0.423671	1.301756	1.147874	2.32626	0	-1.39469	-3.66246	prot kin
degS	0.506373	-0.19513	-0.22035	-0.2241	0.092188	-0.09121	0.38987	0	1.073326	0.208864	-2.28022	prot kin
natK	-0.85093	-0.0638	-0.66019	0.0052	0	0.300397	0.400133	0.279776	0.440304	-1.61448	-4.03108	prot kin
yufL	-0.47825	0.226758	-0.27817	-0.34343	0	0.413216	0.042228	0.229729	0.784527	-0.27551	-2.55484	prot kin
yclK	0.317173	0.133189	0.127189	-0.04495	0	0.137522	-0.31473	0.018506	-0.24772	-1.06706	-3.40539	prot kin
comP	-1.11501	0.113287	-0.20167	0.461646	0.35788	0.542855	0	0.053557	-0.40582	-0.94523	-2.26156	prot kin
resE	1.316994	0	0.788489	0.831102	0.917498	0.807911	-0.0276	-0.18798	-0.01999	-0.90494	-3.23311	prot kin
dctS	0.468179	0.311911	0	-0.30207	0.272696	-0.14286	-0.05766	0.107212	1.014948	-0.35045	-3.56735	prot kin

Appendix 6. – continued.

Gene symbol	LC	6H	12H	1D	2D	3D	5D	7D	14D	1M	2M	Category
glnJ	0.682292	1.515342	0.251194	0.562124	-0.22745	-0.34934	-0.4525	-0.3504	0.258786	0	-1.94862	prot kin
epsB	-2.56909	0.830022	0.252981	1.017717	0	0.054532	0.466027	-1.53603	-1.1776	-2.45636	-5.26136	prot kin
rsbT	-0.4756	0.289568	0	-0.0137	0.494257	-0.11004	0.389553	0.246027	1.253006	-0.43661	-2.18793	prot kin
kinB	-1.69164	-0.02483	-0.56322	0.101281	0	0.028833	0.821041	0.993276	1.321816	-0.72156	-2.91129	prot kin
cssS	0.79851	-0.54717	0.987294	0.134482	-0.07235	0.137095	-0.25202	-0.19453	0.786458	0	-2.00825	prot kin
ykoH	2.806681	0.366781	-0.48143	0.00424	-0.02987	0.029196	0	-0.07913	0.2637	-1.21423	-4.03999	prot kin
kinC	-1.70325	-0.0622	0.209325	0.291741	0	0.049676	0.123563	0.115582	-0.39645	-1.28527	-3.2976	prot kin
yfiJ	0.014145	0.53823	-0.39439	0.366143	0	-0.27	0.151551	-0.28965	0.555706	-1.15564	-3.7452	prot kin
bceS	1.680906	0.065821	-0.29002	-0.17523	0	0.160498	-0.13705	0.312222	0.249864	-0.59539	-2.72099	prot kin
prkA	-1.66631	-4.23296	-4.85069	-4.81424	2.438267	3.499994	3.779536	3.314381	2.587518	0	-2.61585	prot kin
desK	-1.29321	-0.90793	-0.93039	0.484493	0.05255	0.175597	1.173628	1.085471	-0.02472	0	-1.18528	prot kin
kinA	-1.55733	-0.22852	-0.58529	0	0.718456	1.142327	0.693783	1.16163	0.728487	-0.41314	-2.75013	prot kin
yvcQ	-1.13756	1.059465	0.185375	0.241493	0.185305	0.037614	-0.23618	-0.24507	0	-1.33925	-3.44087	prot kin
kinE	-0.29432	-0.07407	-0.26027	0.358893	1.167115	1.224515	0.905698	1.24183	0	-0.52273	-3.3651	prot kin
ywpD	-0.63741	0.731251	-0.32431	0.063173	0	0.305619	0.215498	0.508435	-0.22274	-0.40518	-3.01758	prot kin
ybdK	0.366428	0.248948	0	-0.14588	0.107134	0.080782	-0.13637	-0.17285	0.556576	-0.86981	-3.12525	prot kin
prkC	1.258804	0.785069	-0.30842	0.056043	0.346393	0	-0.0415	0.035133	-0.57989	-1.43223	-3.63777	prot kin
cheA	1.356879	2.225825	2.286033	1.588735	0	0.111444	-0.18542	-0.04394	-0.86707	-0.80869	-2.91685	prot kin
citS	0.626027	-0.37784	-0.40986	-0.26422	0	0.013046	0.149194	0.510582	0.768653	-0.21543	-2.10405	prot kin
phoR	1.425662	-0.68202	-0.82844	-1.21911	0	0.264998	0.282128	0.477733	0.713369	-0.19573	-2.02753	prot kin
hprK	0	0.475441	0.163061	0.311677	0.188306	-0.49223	-0.46329	-0.67938	0.074021	-1.4017	-3.30074	prot kin
lytS	0.034868	0.590561	-0.31559	0.071108	0	0.058358	-0.42359	-0.25018	0.767186	-0.60019	-2.65291	prot kin
ybdM	-0.24312	0.022332	-0.36205	0	0.045411	0.236526	-0.0913	0.150935	0.130628	-0.75689	-2.74668	prot kin
yabT	-0.21128	-2.05265	-2.68448	-2.39878	0	1.144999	0.919721	1.415726	1.054527	0.364485	-1.39215	prot kin
epsA	-5.90862	1.558222	0.922522	1.574576	0.293595	0	0.588095	-2.41714	-1.70694	-2.15491	-7.45872	prot kin
liaS	2.127627	0.31771	-0.27445	-0.36018	0.159595	-0.17159	0.165387	0	1.259787	-0.39163	-2.61869	prot kin
yrkQ	0.383692	-1.60093	-1.89828	-1.97686	-0.33612	0.568199	0	0.432502	1.300322	0.195416	-1.65945	prot kin
ydfH	-1.92611	0.929727	-0.03505	0.200967	0.217494	0	0.128583	0.132496	-0.77494	-0.92786	-3.38921	prot kin
ptkA	-1.87656	1.632419	0.370859	1.116257	0.534015	0	0.067774	-0.18457	-1.15126	-1.35043	-4.67627	prot kin
yesM	0	-2.77746	-3.96048	-3.4314	0.032796	1.119365	0.959075	2.124444	2.302853	-0.00816	-2.2648	prot kin
kinD	0.846196	-0.73045	-0.70931	-0.62638	-0.02213	0.212209	0	0.249837	0.539174	0.289407	-1.72481	prot kin
yxdK	1.780837	0.698052	-0.46635	0.017347	-0.12743	0	-0.11897	0.095001	0.080821	-1.13438	-3.07878	prot kin
yhcY	1.148331	-0.38634	-0.63141	-0.1604	0	0.239996	0.310342	0.394902	0.338793	-0.76887	-3.3916	prot kin
ywqE	0.329975	0.650228	-0.32178	0.865859	0.231282	0	0.281219	-0.07873	-0.3171	-1.39409	-5.44372	prot phosp
cheC	1.914772	1.871262	2.127456	1.399146	-0.37469	-0.43771	-0.45724	0	0.480438	-0.4011	-1.99929	prot phosp
rapJ	-0.29453	0.65759	0.941885	0.978035	0.918646	0.365075	0	-0.13874	-0.98336	-0.76756	-2.24531	prot phosp
rapD	-1.01313	-0.1281	0.189461	-0.0343	0	0.390166	0.061987	0.31232	0.178103	-0.51169	-2.67088	prot phosp
yfkJ	0.176756	-3.51134	-3.296	-4.60754	-0.88215	0.276761	0.161001	0	0.857774	0.226502	-1.48194	prot phosp
rsbP	-1.04478	0.539364	-0.26996	0.222166	0.129577	0.006782	0.023812	0	-0.38969	-0.89766	-2.56801	prot phosp
rsbX	0	-4.38408	-2.48899	-3.79919	-0.4153	1.344889	0.666112	0.771062	1.704487	1.142842	-0.69658	prot phosp
rsbU	-2.04758	0.294858	-0.04424	0.041827	0.310724	-0.31451	0	0.028467	0.186581	-0.41478	-1.60253	prot phosp
spoIIE	1.341061	-1.67423	-2.10258	-2.29421	0	1.266547	0.969837	2.130904	1.12319	-0.95185	-2.72604	prot phosp
rapA	4.93363	-3.07831	-2.89356	-4.34778	-1.8626	0	-0.19159	1.021603	2.452611	2.497477	0.197378	prot phosp
rapH	2.93347	-2.10697	-0.65958	-0.74977	0.253411	0.538349	0	0.395494	0.282748	-0.25986	-2.92229	prot phosp
rapC	-1.21615	2.335645	-1.14712	-1.31909	1.043432	1.59807	1.218877	1.157051	0	-0.58944	-2.95531	prot phosp
spo0E	-0.71619	-1.26895	0.035437	0	1.239192	1.528959	0.759984	1.120419	-0.04091	-0.14013	-1.68604	prot phosp
rapG	1.333812	-0.73357	0.111828	-0.32171	-0.02438	0.186924	0.125283	0.48226	0	-1.02713	-3.57291	prot phosp
yjbP	-1.90716	0	-0.48632	-0.117	0.202209	0.545482	0.519845	0.543513	0.174933	-0.82599	-4.70237	prot phosp
fliY	-0.20597	2.468793	2.217446	2.000885	0.47285	0.46141	-0.19874	0	-0.08416	-0.26485	-2.72228	prot phosp
rapB	-1.45468	0.022459	-0.21582	0.229854	0.338669	0.218853	0	0.01251	-0.87036	-0.68519	-2.91759	prot phosp
rapE	-1.27919	0	-2.10852	-1.23861	-0.06993	1.008483	0.387274	1.166708	0.739897	0.561649	-0.49191	prot phosp
rapK	-3.47755	1.020281	0.042604	1.876368	1.348318	1.11024	-0.00968	0	-3.83322	-1.38625	-2.67977	prot phosp
rapI	-1.30126	0.998837	-0.6248	0.972828	0.600125	0.289668	0.015467	0	-2.75366	-1.79617	-2.46105	prot phosp
prpC	0.899514	0.57057	-0.15333	-0.13368	0.22537	0	0.049266	-0.01848	0.184267	-1.27996	-3.91543	prot phosp
ywlE	-0.26592	0	0.089595	-0.23856	-0.04278	0.18376	0.544381	0.533635	0.301132	-0.49887	-2.83921	prot phosp
yisI	-1.77972	-0.71738	-2.43373	-1.9667	0.653507	1.211429	1.216839	1.073426	0	0.575282	-0.19361	prot phosp
rapF	1.994238	2.676807	-1.48967	-1.08808	-0.06366	1.356047	0.390709	1.150483	0	-0.48624	-1.13493	prot phosp



Appendix 7. Biofilm ontogeny is a punctuated process organized in functionally discrete stages. Enrichment analysis of SubtiWiki functional categories (maximal ontology depth) in a respective biofilm growth timepoint for genes with transcript expression 0.5 times (log_2 scale) above the median of their overall transcription profile. Colouring follows biofilm growth periods: LC (grey), early (red), mid (blue), late (green). Functional enrichment is tested by one-tailed hypergeometric test and *p* values are adjusted for multiple testing (see Material and Methodology). Figure obtained from Futo *et al.* (2021).

Appendix 8. Biofilm ontogeny is a punctuated process organized in functionally discrete stages. Enrichment analysis of SubtiWiki functional categories (ontology depth 1) in a respective biofilm growth timepoint for genes with transcript expression 0.5 times (log_2 scale) above the median of their overall transcription profile. Functional enrichment is tested by one-tailed hypergeometric test and *p* values are adjusted for multiple testing (see Material and Methodology). q= the number of specific annotation at specific timepoint; s = the number of all annotations at a specific timepoint; h = the number of specific annotation in all timepoints; t = the number of all annotations in all timepoints. Table obtained from Futo *et al.* (2021).

LC	**q**	**s**	**h**	**t**	**p**	**padj**	**log_odds**
Metabolism	448	2076	1167	7067	2,59E-13	3,33E-12	0,71
6H	**q**	**s**	**h**	**t**	**p**	**padj**	**log_odds**
Information processing	455	2511	1017	7067	3,69E-11	4,06E-10	0,65
Groups of genes	908	2511	2360	7067	0,000145	0,000655	0,28
Metabolism	459	2511	1167	7067	0,001749	0,007241	0,28
12H	**q**	**s**	**h**	**t**	**p**	**padj**	**log_odds**
Metabolism	323	1609	1167	7067	9,97E-06	5,91E-05	0,46
Prophages and mobile genetic elements	106	1609	338	7067	0,00012	0,000579	0,67
1D	**q**	**s**	**h**	**t**	**p**	**padj**	**log_odds**
no_annotation	58	1704	151	7067	5,47E-05	0,000296	1
Groups of genes	608	1704	2360	7067	0,011884	0,043575	0,19
2D	**q**	**s**	**h**	**t**	**p**	**padj**	**log_odds**
Lifestyles	524	2017	1373	7067	4,59E-18	7,06E-17	0,8
no_annotation	69	2017	151	7067	4,45E-06	2,85E-05	1,1
3D	**q**	**s**	**h**	**t**	**p**	**padj**	**log_odds**
Lifestyles	704	2450	1373	7067	1,79E-45	4,59E-44	1,25
no_annotation	70	2450	151	7067	0,001787	0,007241	0,72
5D	**q**	**s**	**h**	**t**	**p**	**padj**	**log_odds**
Lifestyles	686	2121	1373	7067	5,86E-68	4,51E-66	1,57
7D	**q**	**s**	**h**	**t**	**p**	**padj**	**log_odds**
Lifestyles	724	2509	1373	7067	2,14E-48	8,23E-47	1,29
14D	**q**	**s**	**h**	**t**	**p**	**padj**	**log_odds**
Lifestyles	749	3109	1373	7067	1,37E-18	2,64E-17	0,76
Cellular processes	350	3109	661	7067	7,44E-07	5,20E-06	0,57
Metabolism	550	3109	1167	7067	0,009993	0,038471	0,22
1M	**q**	**s**	**h**	**t**	**p**	**padj**	**log_odds**
Lifestyles	178	648	1373	7067	1,14E-07	8,80E-07	0,73
Prophages and mobile genetic elements	53	648	338	7067	5,77E-05	0,000296	0,94
2M	**q**	**s**	**h**	**t**	**p**	**padj**	**log_odds**
Prophages and mobile genetic elements	26	134	338	7067	6,05E-10	5,83E-09	2,35
Lifestyles	56	134	1373	7067	1,66E-09	1,42E-08	1,61

Appendix 9. Biofilm ontogeny is a punctuated process organized in functionally discrete stages. Enrichment analysis of SubtiWiki functional categories (ontology depth 2) in a respective biofilm growth timepoint for genes with transcript expression 0.5 times (log_2 scale) above the median of their overall transcription profile. Functional enrichment is tested by one-tailed hypergeometric test and *p* values are adjusted for multiple testing (see Material and Methodology). q = the number of specific annotation at specific timepoint; s = the number of all annotations at a specific timepoint; h = the number of specific annotation in all timepoints; t = the number of all annotations in all timepoints. Table obtained from Futo *et al.* (2021).

LC	q	s	h	t	p	padj	log_odds
Metabolism#Carbon metabolism	182	2337	293	7850	1.55E-31	8.88E-30	2.04
Groups of genes#Phosphoproteins	124	2337	324	7850	0.000503	0.004925	0.57
Groups of genes#Universally conserved proteins	17	2337	28	7850	0.000649	0.006185	1.87
Cellular processes#Transporters	148	2337	408	7850	0.002175	0.018193	0.45
Cellular processes#Homeostasis	44	2337	104	7850	0.004189	0.031232	0.8

6H	q	s	h	t	p	padj	log_odds
Groups of genes#Essential genes	179	2829	257	7850	9.09E-29	4.45E-27	2.1
Information processing#Protein synthesis, modification and	232	2829	401	7850	6.07E-20	2.31E-18	1.36
Groups of genes#Universally conserved	28	2829	28	7850	3.57E-13	8.73E-12	NA
Lifestyles#Exponential and early post-exponential lifestyles	118	2829	194	7850	1.08E-12	2.31E-11	1.5
Metabolism#Nucleotide metabolism	72	2829	111	7850	4.97E-10	9.47E-09	1.74
Metabolism#Amino acid/ nitrogen metabolism	117	2829	269	7850	0.006176	0.044133	0.47
Metabolism#Detoxification reactions	9	2829	12	7850	0.00697	0.048788	2.42

12H	q	s	h	t	p	padj	log_odds
Lifestyles#Exponential and early post-exponential lifestyles	89	1779	194	7850	4.23E-13	9.68E-12	1.58
Metabolism#Additional metabolic pathways	125	1779	367	7850	2.18E-07	3.40E-06	0.86
Prophages and mobile genetic elements#Prophages	104	1779	311	7850	5.98E-06	7.88E-05	0.81
Groups of genes#Secreted proteins	47	1779	116	7850	1.11E-05	0.000136	1.24
Metabolism#Amino acid/ nitrogen metabolism	90	1779	269	7850	2.47E-05	0.000274	0.81
Groups of genes#Universally conserved	15	1779	28	7850	0.000358	0.003613	1.99
Groups of genes#Phosphoproteins	98	1779	324	7850	0.000764	0.00708	0.59
Metabolism#Electron transport and ATP	32	1779	91	7850	0.004323	0.031552	0.9

Appendix 9. – continued.

1D	q	s	h	t	p	padj	log_odds
Lifestyles#Exponential and early post-exponential lifestyles	102	1910	194	7850	9.51E-18	2.96E-16	1.84
Metabolism#Nucleotide metabolism	51	1910	111	7850	4.72E-07	7.04E-06	1.43
Cellular processes#Homeostasis	47	1910	104	7850	2.36E-06	3.37E-05	1.38
no_annotation	58	1910	151	7850	7.32E-05	0.000785	0.98
Groups of genes#Proteins of unknown	205	1910	708	7850	0.00178	0.015263	0.38
Metabolism#Detoxification reactions	8	1910	12	7850	0.002289	0.018696	2.64
Groups of genes#Secreted proteins	42	1910	116	7850	0.00265	0.021141	0.83
Metabolism#Electron transport and ATP	34	1910	91	7850	0.003652	0.027834	0.9

2D	q	s	h	t	p	padj	log_odds
Lifestyles#Sporulation	372	2219	647	7850	1.08E-59	7.43E-58	1.97
no_annotation	69	2219	151	7850	3.15E-06	4.32E-05	1.12
Cellular processes#Homeostasis	44	2219	104	7850	0.001396	0.012273	0.91

3D	q	s	h	t	p	padj	log_odds
Lifestyles#Sporulation	513	2674	647	7850	1.26E-134	1.44E-132	3.16
no_annotation	70	2674	151	7850	0.001063	0.009596	0.76
Metabolism#Electron transport and ATP	44	2674	91	7850	0.003224	0.025133	0.87

5D	q	s	h	t	p	padj	log_odds
Lifestyles#Sporulation	523	2288	647	7850	3.17E-180	1.09E-177	3.7

7D	q	s	h	t	p	padj	log_odds
Lifestyles#Sporulation	523	2730	647	7850	1.03E-139	1.77E-137	3.26
Metabolism#Carbon metabolism	153	2730	293	7850	3.38E-10	6.82E-09	1.08
Groups of genes#Efp-dependent prot	22	2730	34	7850	0.000349	0.003613	1.79

14D	q	s	h	t	p	padj	log_odds
Lifestyles#Sporulation	488	3421	647	7850	3.91E-66	3.35E-64	2.16
Metabolism#Carbon metabolism	209	3421	293	7850	1.18E-22	5.05E-21	1.75
Cellular processes#Transporters	252	3421	408	7850	2.64E-14	6.97E-13	1.12
Groups of genes#Membrane proteins	558	3421	1076	7850	2.61E-09	4.71E-08	0.56
Groups of genes#Efp-dependent prot	27	3421	34	7850	2.11E-05	0.000241	2.33

1M	q	s	h	t	p	padj	log_odds
Lifestyles#Coping with stress	113	708	603	7850	5.37E-15	1.54E-13	1.37
Metabolism#Carbon metabolism	55	708	293	7850	8.56E-08	1.40E-06	1.29
Prophages and mobile genetic elements#Prophages	51	708	311	7850	1.58E-05	0.000186	1.04

Appendix 9. – continued.

2M	q	s	h	t	p	padj	log_odds
Lifestyles#Sporulation	49	141	647	7850	2.55E-19	8.75E-18	2.66
Prophages and mobile genetic elements#Mobile genetic elements	8	141	27	7850	1.48E-08	2.53E-07	4.61
Prophages and mobile genetic elements#Prophages	18	141	311	7850	1.03E-05	0.000131	1.89

Appendix 10. Biofilm ontogeny is a punctuated process organized in functionally discrete stages. Enrichment analysis of SubtiWiki functional categories (ontology depth 4) in a respective biofilm growth timepoint for genes with transcript expression 0.5 times (log_2 scale) above the median of their overall transcription profile. Functional enrichment is tested by one-tailed hypergeometric test and *p* values are adjusted for multiple testing (see Material and Methodology). q = the number of specific annotation at specific timepoint; s = the number of all annotations at a specific timepoint; h = the number of specific annotation in all timepoints; t = the number of all annotations in all timepoints. Table obtained from Futo *et al.* (2021).

LC	q	s	h	t	p	padj	log_odds
Metabolism#Carbon metabolism# Utilization of specific carbon sources# Utilization of pectin	18	2513	19	8374	5.09E-09	3.73E-07	5.4
Metabolism#Lipid metabolism# Utilization of lipids#Utilization of fatty acids	15	2513	19	8374	1.46E-05	0.000494	3.14
Metabolism#Carbon metabolism# Utilization of specific carbon sources# Utilization of starch/ maltodextrin	11	2513	12	8374	1.52E-05	0.000509	4.69
Metabolism#Carbon metabolism# Utilization of specific carbon sources# Utilization of glucomanNA	8	2513	8	8374	6.53E-05	0.001711	NA
Metabolism#Nucleotide metabolism# Utilization of nucleotides	15	2513	21	8374	0.000107	0.002629	2.55
Metabolism#Carbon metabolism# Utilization of specific carbon sources# Utilization of arabiNA/ arabinose/ arabitol	10	2513	12	8374	0.000205	0.004446	3.55
Metabolism#Carbon metabolism# Utilization of specific carbon sources# Utilization of inositol	13	2513	18	8374	0.000266	0.005422	2.61
Information processing#Protein synthesis, modification and degradation#Translation# Ribosomal proteins	30	2513	57	8374	0.000285	0.005602	1.38
Cellular processes#Transporters# Transporters/ other#Metal ion transporter	18	2513	29	8374	0.000341	0.006448	1.94
Metabolism#Additional metabolic pathways# Biosynthesis of cofactors#Biosynthesis/ acquisition of thiamine	13	2513	19	8374	0.000611	0.010673	2.34
Groups of genes#Universally conserved protei	17	2513	28	8374	0.00072	0.011897	1.86

Appendix 10. – continued.

Metabolism#Carbon metabolism# Utilization of specific carbon sources# Utilization of glucarate/galactarate	6	2513	6	8374	0.000727	0.011897	NA
Metabolism#Additional metabolic pathways# Iron metabolism#Acquisition of iron / Other	13	2513	20	8374	0.001267	0.01778	2.12
Metabolism#Carbon metabolism# Utilization of specific carbon sources# Utilization of rhamnose	5	2513	5	8374	0.002427	0.030268	NA
Information processing#Genetics#DNA repair/ recombination#Other proteins	26	2513	53	8374	0.002688	0.032865	1.18
Metabolism#Amino acid/ nitrogen metabolism#Utilization of nitrogen sources other than amino acids#Utilization of peptides	12	2513	19	8374	0.002802	0.033167	2
Cellular processes#Transporters#ABC transporters# Importers	49	2513	116	8374	0.003247	0.037259	0.78
Information processing#Genetics#DNA repair/ recombination	28	2513	59	8374	0.003461	0.039519	1.08
Metabolism#Carbon metabolism# Utilization of specific carbon sources# Utilization of other polymeric carbohydrates	14	2513	24	8374	0.0036	0.040874	1.71
Metabolism#Carbon metabolism# Utilization of specific carbon sources# Utilization of galactan	6	2513	7	8374	0.003784	0.041992	3.81

6H	**q**	**s**	**h**	**t**	**p**	**padj**	**log_odds**
Groups of genes#Essential genes	179	3022	257	8374	1.27E-28	2.50E-26	2.09
Information processing#Protein synthesis, modification and degradation#Translation# Ribosomal proteins	53	3022	57	8374	1.82E-19	2.75E-17	4.58
Lifestyles#Exponential and early post-exponential lifestyles#Motility and chemotaxis#Flagellar proteins	35	3022	37	8374	8.02E-14	9.37E-12	4.97
Groups of genes#Universally conserved proteins	28	3022	28	8374	3.73E-13	3.99E-11	NA
Prophages and mobile genetic elements# Prophages#PBSX prophage	35	3022	41	8374	9.80E-11	8.99E-09	3.38

Appendix 10. – continued.

Information processing#Protein synthesis, modification and degradation#Translation# Aminoacyl-tRNA synthetases	25	3022	28	8374	7.42E-09	5.02E-07	3.89
Lifestyles#Exponential and early post-exponential lifestyles#Biofilm formation#Matrix polysaccharide synthesis	16	3022	16	8374	8.07E-08	4.51E-06	NA
Metabolism#Nucleotide metabolism# Biosynthesis/ acquisition of nucleotides# Biosynthesis/ acquisition of purine nucleotides	24	3022	28	8374	8.50E-08	4.64E-06	3.42
Metabolism#Amino acid/ nitrogen metabolism# Biosynthesis/ acquisition of amino acids#Biosynthesis/	18	3022	20	8374	8.63E-07	4.10E-05	4
Information processing#Protein synthesis, modification and degradation#Translation#tRNA modification	30	3022	42	8374	3.24E-06	0.000134	2.16
Lifestyles#Exponential and early post-exponential lifestyles#Motility and chemotaxis#Signal transduction	25	3022	33	8374	3.83E-06	0.000156	2.48
Metabolism#Carbon metabolism#Carbon core metabolism#TCA cycle	16	3022	18	8374	5.41E-06	0.000214	3.83
Metabolism#Nucleotide metabolism# Biosynthesis/ acquisition of nucleotides# Biosynthesis/ acquisition of pyrimidine nucleotides	17	3022	20	8374	9.55E-06	0.000346	3.33
Information processing#Genetics# Newly identified competence genes	19	3022	25	8374	5.47E-05	0.00148	2.5
Information processing#Protein synthesis, modification and degradation#Chaperones/ protein folding	11	3022	12	8374	0.000107	0.002629	4.29
Cellular processes#Transporters#Transporters/ other#Nucleotide/ nucleoside transporter	10	3022	11	8374	0.000275	0.005511	4.15
Metabolism#Additional metabolic pathways#Biosynthesis of cofactors#Biosynthesis/ acquisition of biotin	8	3022	8	8374	0.000286	0.005602	NA
Lifestyles#Exponential and early post-exponential lifestyles# Biofilm formation#Other proteins required for efficient	9	3022	10	8374	0.000696	0.011691	4
Information processing#Genetics#DNA condensation/ segregation	13	3022	17	8374	0.000816	0.012788	2.53

Appendix 10. – continued.

Metabolism#Nucleotide metabolism#Utilization of nucleotides	15	3022	21	8374	0.001047	0.015501	2.15
Lifestyles#Exponential and early post-exponential lifestyles#Swarming	12	3022	16	8374	0.001756	0.023257	2.41
Cellular processes#Cell envelope and cell division# Cell wall synthesis#Biosynthesis of teichoic acid	14	3022	20	8374	0.002111	0.02711	2.05
Metabolism#Additional metabolic pathways# Biosynthesis of cell wall components#Biosynthesis of teichoic acid	14	3022	20	8374	0.002111	0.02711	2.05
Metabolism#Additional metabolic pathways#Sulfur metabolism#Conversion of S-(2-succino)cysteine to cysteine	6	3022	6	8374	0.002202	0.027884	NA
Information processing#Protein synthesis, modification and degradation#Translation#Translation factors	9	3022	11	8374	0.002594	0.032038	3
Metabolism#Amino acid/ nitrogen metabolism#Biosynthesis/ acquisition of amino acids#Biosynthesis/ acquisition of histidine	9	3022	11	8374	0.002594	0.032038	3
Metabolism#Carbon metabolism#Carbon core metabolism# Glycolysis	10	3022	13	8374	0.003249	0.037259	2.57
Groups of genes#Phosphoproteins#Phosphorylation on an Arg residue	55	3022	113	8374	0.003862	0.042582	0.76
Lifestyles#Exponential and early post-exponential lifestyles# Biofilm formation	29	3022	53	8374	0.004188	0.045012	1.11
Groups of genes#Phosphoproteins#Phosphorylation on either a Ser, Thr or Tyr residue	23	3022	40	8374	0.004581	0.048233	1.27

12H	**q**	**s**	**h**	**t**	**p**	**padj**	**log_odds**
Prophages and mobile genetic elements#Prophages#PBSX prophage	38	1915	41	8374	1.71E-21	2.92E-19	5.45
Lifestyles#Exponential and early post-exponential lifestyles# Motility and chemotaxis#Flagellar proteins	32	1915	37	8374	3.25E-16	4.28E-14	4.46

Appendix 10. – continued.

Metabolism#Amino acid/ nitrogen metabolism#Biosynthesis/ acquisition of amino acids#Biosynthesis/ acquisition of cysteine	16	1915	16	8374	5.33E-11	5.07E-09	NA
Information processing#Protein synthesis, modification and degradation#Translation#Ribosomal proteins	33	1915	57	8374	1.17E-08	7.50E-07	2.23
Lifestyles#Exponential and early post-exponential lifestyles# Biofilm formation#Matrix polysaccharide synthesis	14	1915	16	8374	7.68E-08	4.39E-06	4.57
Lifestyles#Exponential and early post-exponential lifestyles# Motility and chemotaxis#Signal transduction in motility and	22	1915	33	8374	9.66E-08	5.17E-06	2.77
Metabolism#Nucleotide metabolism#Biosynthesis/ acquisition of nucleotides#Biosynthesis/ acquisition of purine nucleotides	19	1915	28	8374	4.88E-07	2.36E-05	2.84
Metabolism#Amino acid/ nitrogen metabolism#Biosynthesis/ acquisition of amino acids#Biosynthesis/ acquisition of methionine/ S-	15	1915	20	8374	1.10E-06	4.87E-05	3.35
Lifestyles#Exponential and early post-exponential lifestyles# Biofilm formation#Other proteins required for efficient pellicle	9	1915	10	8374	1.34E-05	0.000478	4.93
Groups of genes#Secreted proteins	47	1915	116	8374	1.44E-05	0.000494	1.22
Metabolism#Electron transport and ATP synthesis# Respiration#Anaerobic respiration	8	1915	9	8374	5.31E-05	0.001451	4.76
Metabolism#Additional metabolic pathways#Sulfur metabolism#sulfur metabolism/ general	11	1915	16	8374	0.000119	0.002887	2.9
Metabolism#Additional metabolic pathways#Sulfur metabolism#Conversion of S-(2-succino)cysteine to cysteine	6	1915	6	8374	0.000142	0.003351	NA
Groups of genes#Phosphoproteins#Phosphorylation on either a Ser, Thr or Tyr residue	20	1915	40	8374	0.000155	0.003622	1.76
Metabolism#Additional metabolic pathways#Biosynthesis of cofactors#Biosynthesis/ acquisition of biotin	7	1915	8	8374	0.000208	0.004446	4.57
Metabolism#Additional metabolic pathways#Sulfur metabolism#Conversion of S-methyl cysteine to cysteine	7	1915	8	8374	0.000208	0.004446	4.57

Appendix 10. – continued.

Metabolism#Additional metabolic pathways#Sulfur metabolism	10	1915	15	8374	0.000364	0.00677	2.76
Cellular processes#Homeostasis#Acquisition of iron#Acquisition of iron / Other	12	1915	20	8374	0.000385	0.007114	2.35
Groups of genes#Universally conserved proteins	15	1915	28	8374	0.000398	0.00731	1.97
Metabolism#Carbon metabolism#Carbon core metabolism# Glycolysis	9	1915	13	8374	0.000483	0.008674	2.93
Metabolism#Amino acid/ nitrogen metabolism#Biosynthesis/ acquisition of amino acids#Biosynthesis/ acquisition of histidine	8	1915	11	8374	0.00062	0.010765	3.17
Lifestyles#Exponential and early post-exponential lifestyles# Biofilm formation	23	1915	53	8374	0.000706	0.01178	1.38
Cellular processes#Cell envelope and cell division#Cell wall degradation/ turnover#Autolysis	13	1915	24	8374	0.000849	0.013064	2
Lifestyles#Coping with stress#Biosynthesis of antibacterial compounds	22	1915	51	8374	0.00101	0.015175	1.37
Metabolism#Additional metabolic pathways#Miscellaneous metabolic pathways#Biosynthesis of antibacterial compounds	22	1915	51	8374	0.00101	0.015175	1.37
Information processing#Genetics#Newly identified competence genes	13	1915	25	8374	0.001402	0.019464	1.88
Information processing#Protein synthesis, modification and degradation#Chaperones/ protein folding	8	1915	12	8374	0.001487	0.02043	2.76
Metabolism#Additional metabolic pathways#Biosynthesis of cofactors#Biosynthesis of NAD(P)	8	1915	12	8374	0.001487	0.02043	2.76
Groups of genes#Phosphoproteins#Phosphorylation on an Arg residue	40	1915	113	8374	0.001596	0.021633	0.9
Metabolism#Additional metabolic pathways#Iron metabolism# Acquisition of iron / Other	11	1915	20	8374	0.00182	0.023853	2.05

Appendix 10. – continued.

Metabolism#Amino acid/ nitrogen metabolism#Biosynthesis/ acquisition of amino acids#Biosynthesis/ acquisition of aromatic amino acids	11	1915	20	8374	0.00182	0.023853	2.05
Metabolism#Electron transport and ATP synthesis#ATP synthesis#Substrate-level phosphorylation	4	1915	4	8374	0.002728	0.032865	NA

1D	**q**	**s**	**h**	**t**	**p**	**padj**	**log_odds**
Lifestyles#Exponential and early post-exponential lifestyles# Motility and chemotaxis#Flagellar proteins	32	2054	37	8374	2.80E-15	3.43E-13	4.32
Lifestyles#Exponential and early post-exponential lifestyles# Motility and chemotaxis#Signal transduction in motility and chemotaxis	24	2054	33	8374	7.20E-09	5.00E-07	3.05
Lifestyles#Exponential and early post-exponential lifestyles# Biofilm formation#Matrix polysaccharide synthesis	15	2054	16	8374	8.31E-09	5.47E-07	5.54
Lifestyles#Exponential and early post-exponential lifestyles# Biofilm formation	32	2054	53	8374	2.85E-08	1.74E-06	2.25
Metabolism#Amino acid/ nitrogen metabolism#Biosynthesis/ acquisition of amino acids#Biosynthesis/ acquisition of cysteine	14	2054	16	8374	1.98E-07	1.01E-05	4.44
Cellular processes#Homeostasis#Acquisition of iron#ABC transporters for the uptake of iron/ siderophores	16	2054	21	8374	9.10E-07	4.18E-05	3.31
Metabolism#Additional metabolic pathways#Iron metabolism#ABC transporters for the uptake of iron/ siderophores	16	2054	21	8374	9.10E-07	4.18E-05	3.31
Metabolism#Amino acid/ nitrogen metabolism# Biosynthesis/ acquisition of amino acids# Biosynthesis/ acquisition of methionine/	15	2054	20	8374	2.86E-06	0.00012	3.22
Metabolism#Nucleotide metabolism#Biosynthesis/ acquisition of nucleotides#Biosynthesis/ acquisition	14	2054	20	8374	2.28E-05	0.000698	2.85
Lifestyles#Exponential and early post-exponential lifestyles#Biofilm formation#Other proteins required for efficient pellicle biofilm	9	2054	10	8374	2.47E-05	0.000748	4.8
Lifestyles#Coping with stress#Toxins, antitoxins and immunity against toxins/ based on similarity	13	2054	18	8374	2.69E-05	0.000794	3.01
Metabolism#Nucleotide metabolism#Biosynthesis/ acquisition of nucleotides#Biosynthesis/ acquisition of purine nucleotides	17	2054	28	8374	4.91E-05	0.001356	2.26

Appendix 10. – continued.

no_annotation	58	2054	151	8374	9.40E-05	0.002392	0.96
Information processing#RNA synthesis and degradation#RNA chaperones	6	2054	6	8374	0.000217	0.004561	NA
Metabolism#Additional metabolic pathways#Sulfur metabolism#Conversion of S-(2-succino)cysteine to cysteine	6	2054	6	8374	0.000217	0.004561	NA
Metabolism#Additional metabolic pathways#Biosynthesis of cofactors#Biosynthesis/ acquisition of biotin	7	2054	8	8374	0.000333	0.006342	4.43
Metabolism#Additional metabolic pathways#Sulfur metabolism#Conversion of S-methyl cysteine to cysteine	7	2054	8	8374	0.000333	0.006342	4.43
Cellular processes#Cell envelope and cell division#Cell wall synthesis#Biosynthesis of teichoic acid	12	2054	20	8374	0.000766	0.012079	2.21
Cellular processes#Homeostasis#Acquisition of iron#Acquisition of iron / Other	12	2054	20	8374	0.000766	0.012079	2.21
Metabolism#Additional metabolic pathways#Biosynthesis of cell wall components#Biosynthesis of teichoic acid	12	2054	20	8374	0.000766	0.012079	2.21
Metabolism#Additional metabolic pathways #Iron metabolism#Acquisition of iron / Other	12	2054	20	8374	0.000766	0.012079	2.21
Metabolism#Amino acid/ nitrogen metabolism# Biosynthesis/ acquisition of amino acids# Biosynthesis/ acquisition of histidine	8	2054	11	8374	0.001029	0.015362	3.04
Metabolism#Electron transport and ATP synthesis#Respiration# Anaerobic respiration	7	2054	9	8374	0.00118	0.016803	3.43
Lifestyles#Coping with stress#Cold stress prot	10	2054	16	8374	0.001389	0.019397	2.36
Information processing#Protein synthesis, modification and degradation# Chaperones/ protein folding	8	2054	12	8374	0.002422	0.030268	2.63
Metabolism#Detoxification reactions	8	2054	12	8374	0.002422	0.030268	2.63

Appendix 10. – continued.

Groups of genes#Proteins of unknown functio	205	2054	708	8374	0.002761	0.032865	0.36
Groups of genes#Secreted proteins	42	2054	116	8374	0.003123	0.036298	0.82
Prophages and mobile genetic elements# Prophages#SP-beta prophage	61	2054	181	8374	0.003158	0.03654	0.66
Metabolism#Electron transport and ATP synthesis#ATP synthesis# Substrate-level phosphorylation	4	2054	4	8374	0.003612	0.040874	NA
Groups of genes#Phosphoproteins# Phosphorylation on either a Ser, Thr or Tyr residue	18	2054	40	8374	0.003661	0.041251	1.34
Cellular processes#Homeostasis# Trace metal homeostasis (Cu, Zn, Ni, Mn, Mo)#Zinc	6	2054	8	8374	0.003792	0.041992	3.21
Information processing#Regulation of gene expression#phosphorelay#Proteins controlling the activity of the kinases	6	2054	8	8374	0.003792	0.041992	3.21
Lifestyles#Sporulation#phosphorelay# Proteins controlling the activity of the kinases	6	2054	8	8374	0.003792	0.041992	3.21
Metabolism#Amino acid/ nitrogen metabolism#Utilization of nitrogen sources other than amino acids#Utilization of nitrate/ nitrite	5	2054	6	8374	0.004225	0.045221	3.95

2D	**q**	**s**	**h**	**t**	**p**	**padj**	**log_odds**
Lifestyles#Sporulation#Sporulation proteins#Sporulation proteins/ other	176	2350	274	8374	8.76E-37	2.25E-34	2.29
Lifestyles#Sporulation#Sporulation proteins#Spore coat proteins	60	2350	73	8374	6.34E-22	1.16E-19	3.6
Lifestyles#Sporulation#Sporulation proteins# Newly identified sporulation proteins (based on transcription profiling)	87	2350	173	8374	3.60E-10	3.03E-08	1.41
Lifestyles#Coping with stress#Toxins, antitoxins and immunity against toxins	15	2350	17	8374	3.78E-07	1.91E-05	4.27
no_annotation	69	2350	151	8374	2.41E-06	0.000103	1.13

Appendix 10. – continued.

Cellular processes#Homeostasis#Acquisition of iron#ABC transporters for the uptake of iron/ siderophores	16	2350	21	8374	6.34E-06	0.00024	3.04
Metabolism#Additional metabolic pathways#Iron metabolism#ABC transporters for the uptake of iron/ siderophores	16	2350	21	8374	6.34E-06	0.00024	3.04
Cellular processes#Homeostasis#Acquisition of iron#Acquisition of iron / Other	15	2350	20	8374	1.74E-05	0.000549	2.95
Metabolism#Additional metabolic pathways#Iron metabolism#Acquisition of iron / Other	15	2350	20	8374	1.74E-05	0.000549	2.95
Metabolism#Amino acid/ nitrogen metabolism#Biosynthesis/ acquisition of amino acids#Biosynthesis/ acquisition of methionine/ S-adenosylmethionine	15	2350	20	8374	1.74E-05	0.000549	2.95
Lifestyles#Coping with stress#Toxins, antitoxins and immunity against toxins#Type 2 TA systems	12	2350	16	8374	0.000129	0.003108	2.95
Lifestyles#Sporulation#Germination#Additional germination proteins	18	2350	30	8374	0.000247	0.005122	1.95
Metabolism#Electron transport and ATP synthesis#ATP synthesis#ATPase	8	2350	9	8374	0.000258	0.005299	4.36
Information processing#RNA synthesis and degradation#RNA chaperones	6	2350	6	8374	0.000486	0.008674	NA
Lifestyles#Coping with stress#Toxins, antitoxins and immunity against toxins/ based on similarity	12	2350	18	8374	0.000732	0.011897	2.36
Prophages and mobile genetic elements# Prophages#Prophage 1	12	2350	18	8374	0.000732	0.011897	2.36
Metabolism#Additional metabolic pathways#Biosynthesis of cofactors# Biosynthesis/ acquisition of biotin	7	2350	8	8374	0.000823	0.012807	4.17
Information processing#Protein synthesis, modification and degradation#Proteolysis# Proteolysis during sporulation/ germination	5	2350	5	8374	0.001735	0.023097	NA

3D	**q**	**s**	**h**	**t**	**p**	**padj**	**log_odds**
Lifestyles#Sporulation#Sporulation proteins# Sporulation proteins/ other	230	2838	274	8374	1.11E-67	9.50E-65	3.46

Appendix 10. – continued.

Lifestyles#Sporulation#Sporulation proteins# Newly identified sporulation proteins (based on transcription profiling)	140	2838	173	8374	1.13E-37	3.23E-35	3.11
Lifestyles#Sporulation#Sporulation proteins#Spore coat proteins	70	2838	73	8374	1.36E-29	2.91E-27	5.54
Lifestyles#Sporulation#Germination#Additional germination proteins	27	2838	30	8374	2.36E-10	2.09E-08	4.15
Lifestyles#Coping with stress#General stress proteins (controlled by SigB)	92	2838	159	8374	3.66E-10	3.03E-08	1.45
Lifestyles#Coping with stress#Toxins, antitoxins and immunity against toxins	15	2838	17	8374	5.54E-06	0.000216	3.88
Metabolism#Electron transport and ATP synthesis# Respiration#Terminal oxidases	13	2838	16	8374	0.000139	0.0033	3.09
Cellular processes#Cell envelope and cell division# Cell wall synthesis#Biosynthesis of teichuronic acid	8	2838	8	8374	0.000173	0.003896	NA
Lifestyles#Sporulation#Germination/ based on similarity	8	2838	8	8374	0.000173	0.003896	NA
Metabolism#Additional metabolic pathways# Biosynthesis of cell wall components# Biosynthesis of teichuronic acid	8	2838	8	8374	0.000173	0.003896	NA
Lifestyles#Sporulation#Germination# GermiNAt receptors	9	2838	10	8374	0.000407	0.007413	4.14
Prophages and mobile genetic elements# Prophages#SP-beta prophage	82	2838	181	8374	0.000842	0.013026	0.71
no_annotation	70	2838	151	8374	0.000915	0.013991	0.77
Metabolism#Electron transport and ATP synthesis#Respiration#Respiration/ other	9	2838	11	8374	0.00156	0.021311	3.14
Metabolism#Additional metabolic pathways# Biosynthesis of cofactors#Biosynthesis/ acquisition of biotin	7	2838	8	8374	0.002878	0.033758	3.77
Information processing#Regulation of gene expression#Quorum sensing	14	2838	22	8374	0.004101	0.04427	1.78

Appendix 10. – continued.

Information processing#Protein synthesis, modification and degradation#Proteolysis# Proteolysis during sporulation/ germination	5	2838	5	8374	0.004461	0.047351	NA
Metabolism#Additional metabolic pathways# Miscellaneous metabolic pathways#Biosynthesis of glycogen	5	2838	5	8374	0.004461	0.047351	NA
Prophages and mobile genetic elements# Prophages#Prophage 1	12	2838	18	8374	0.004538	0.047974	1.97

5D	**q**	**s**	**h**	**t**	**p**	**padj**	**log_odds**
Lifestyles#Sporulation#Sporulation proteins# Sporulation proteins/ other	237	2408	274	8374	8.75E-92	2.25E-88	4.13
Lifestyles#Sporulation#Sporulation proteins#Newly identified sporulation proteins (based on transcription profiling)	146	2408	173	8374	1.96E-53	1.26E-50	3.83
Lifestyles#Sporulation#Sporulation proteins# Spore coat proteins	72	2408	73	8374	2.62E-38	8.89E-36	7.52
Lifestyles#Sporulation#Germination#Additional germination proteins	28	2408	30	8374	1.43E-13	1.59E-11	5.13
Cellular processes#Cell envelope and cell division# Cell wall synthesis#Biosynthesis of teichuronic acid	8	2408	8	8374	4.64E-05	0.001295	NA
Metabolism#Additional metabolic pathways# Biosynthesis of cell wall components# Biosynthesis of teichuronic acid	8	2408	8	8374	4.64E-05	0.001295	NA
Lifestyles#Coping with stress#Toxins, antitoxins and immunity against toxins	13	2408	17	8374	6.23E-05	0.00165	3.02
Lifestyles#Sporulation#Germination# GermiNAt receptors	9	2408	10	8374	9.87E-05	0.002461	4.48
Lifestyles#Coping with stress#Cold stress prot	12	2408	16	8374	0.000168	0.003882	2.9
Metabolism#Amino acid/ nitrogen metabolism# Biosynthesis/ acquisition of amino acids# Biosynthesis/ acquisition of arginine	11	2408	16	8374	0.00105	0.015501	2.45
Information processing#Protein synthesis, modification and degradation#Proteolysis# Proteolysis during sporulation/ germination	5	2408	5	8374	0.00196	0.025435	NA

Appendix 10. – continued.

Metabolism#Additional metabolic pathways# Miscellaneous metabolic pathways# Biosynthesis of glycogen	5	2408	5	8374	0.00196	0.025435	NA
Metabolism#Additional metabolic pathways# Phosphate metabolism	11	2408	17	8374	0.002203	0.027884	2.19
Metabolism#Amino acid/ nitrogen metabolism# Utilization of amino acids#Utilization of branched-chain amino acids	13	2408	22	8374	0.002834	0.033396	1.85

7D	**q**	**s**	**h**	**t**	**p**	**padj**	**log_odds**
Lifestyles#Sporulation#Sporulation proteins# Sporulation proteins/ other	237	2868	274	8374	3.25E-74	4.17E-71	3.73
Lifestyles#Sporulation#Sporulation proteins# Newly identified sporulation proteins (based on transcription profiling)	143	2868	173	8374	5.96E-40	3.06E-37	3.26
Lifestyles#Sporulation#Sporulation proteins# Spore coat proteins	72	2868	73	8374	8.42E-33	1.97E-30	7.15
Lifestyles#Sporulation#Germination# Additional germination proteins	28	2868	30	8374	1.68E-11	1.66E-09	4.76
Metabolism#Carbon metabolism#Utilization of specific carbon sources#Utilization of pectin	19	2868	19	8374	1.38E-09	1.06E-07	NA
Prophages and mobile genetic elements# Prophages#PBSX prophage	28	2868	41	8374	8.68E-06	0.000319	2.06
Metabolism#Carbon metabolism#Utilization of specific carbon sources#Utilization of hexuronate	13	2868	15	8374	4.30E-05	0.00124	3.65
Metabolism#Amino acid/ nitrogen metabolism# Utilization of amino acids#Utilization of branched-chain amino acids	17	2868	22	8374	4.51E-05	0.001288	2.71
Cellular processes#Cell envelope and cell division# Cell wall synthesis#Biosynthesis of teichuronic acid	8	2868	8	8374	0.000188	0.00413	NA
Metabolism#Additional metabolic pathways# Biosynthesis of cell wall components# Biosynthesis of teichuronic acid	8	2868	8	8374	0.000188	0.00413	NA
Metabolism#Carbon metabolism#Utilization of specific carbon sources#Utilization of glucomanNA	8	2868	8	8374	0.000188	0.00413	NA

Appendix 10. – continued.

Groups of genes#Efp-dependent proteins	22	2868	34	8374	0.000273	0.005511	1.82
Lifestyles#Sporulation#Germination# GermiNAt receptors	9	2868	10	8374	0.000445	0.008055	4.12
Metabolism#Carbon metabolism#Utilization of specific carbon sources#Utilization of arabiNA/ arabinose/ arabitol	10	2868	12	8374	0.00069	0.011669	3.27
Metabolism#Amino acid/ nitrogen metabolism# Utilization of nitrogen sources other than amino	8	2868	9	8374	0.001178	0.016803	3.94
Metabolism#Carbon metabolism#Utilization of specific carbon sources#Utilization of melibiose	6	2868	6	8374	0.001608	0.021633	NA
Metabolism#Carbon metabolism#Utilization of specific carbon sources#Utilization of ribose	6	2868	6	8374	0.001608	0.021633	NA
Information processing#Regulation of gene expression#Sigma factors and their control# Control of sigma factors	25	2868	46	8374	0.003942	0.042906	1.2
Information processing#Protein synthesis, modification and degradation#Proteolysis# Proteolysis during sporulation/ germination	5	2868	5	8374	0.004701	0.049098	NA
Metabolism#Additional metabolic pathways#Miscellaneous metabolic pathways#Biosynthesis of glycogen	5	2868	5	8374	0.004701	0.049098	NA
Metabolism#Carbon metabolism#Utilization of specific carbon sources#Utilization of starch/ maltodextrin	9	2868	12	8374	0.004724	0.049133	2.53

14D	**q**	**s**	**h**	**t**	**p**	**padj**	**log_odds**
Lifestyles#Sporulation#Sporulation proteins# Sporulation proteins/ other	221	3611	274	8374	2.77E-38	8.89E-36	2.53
Lifestyles#Sporulation#Sporulation proteins# Spore coat proteins	68	3611	73	8374	9.87E-20	1.59E-17	4.19
Lifestyles#Sporulation#Sporulation proteins#Newly identified sporulation proteins (based on transcription profiling)	127	3611	173	8374	3.33E-16	4.28E-14	1.9
Lifestyles#Coping with stress#General stress proteins (controlled by SigB)	111	3611	159	8374	6.19E-12	6.36E-10	1.64

Appendix 10. – continued.

Groups of genes#Membrane proteins	558	3611	1076	8374	4.17E-10	3.35E-08	0.58
Cellular processes#Transporters# ABC transporters#Importers	82	3611	116	8374	1.40E-09	1.06E-07	1.69
Metabolism#Carbon metabolism#Utilization of specific carbon sources#Utilization of pectin	19	3611	19	8374	1.12E-07	5.85E-06	NA
Lifestyles#Sporulation#Germination# Additional germination proteins	26	3611	30	8374	9.82E-07	4.43E-05	3.11
Prophages and mobile genetic elements# Prophages#PBSX prophage	32	3611	41	8374	5.31E-06	0.000213	2.24
Cellular processes#Transporters# Phosphotransferase system#Sugar specific PTS proteins	21	3611	24	8374	8.58E-06	0.000319	3.21
Metabolism#Amino acid/ nitrogen metabolism#Utilization of nitrogen sources other than amino acids#Utilization of amino sugars	16	3611	17	8374	1.42E-05	0.000493	4.41
Metabolism#Carbon metabolism#Utilization of specific carbon sources#Utilization of amino sugars	16	3611	17	8374	1.42E-05	0.000493	4.41
Groups of genes#Efp-dependent proteins	27	3611	34	8374	1.68E-05	0.000549	2.36
Metabolism#Carbon metabolism#Utilization of specific carbon sources#Utilization of organic acids	25	3611	32	8374	5.77E-05	0.001543	2.24
Groups of genes#Phosphoproteins#Phosphorylation on a Cys residue	14	3611	15	8374	6.80E-05	0.001747	4.21
Metabolism#Carbon metabolism#Utilization of specific carbon sources#Utilization of hexuronate	14	3611	15	8374	6.80E-05	0.001747	4.21
Cellular processes#Cell envelope and cell division# Cell wall degradation/ turnover# Utilization of cell wall components	11	3611	11	8374	9.50E-05	0.002394	NA
Lifestyles#Sporulation#Germination# GermiNAt receptors	10	3611	10	8374	0.000221	0.00461	NA

Appendix 10. – continued.

Metabolism#Lipid metabolism#Utilization of lipids#Utilization of fatty acids	16	3611	19	8374	0.000288	0.005602	2.82
Cellular processes#Transporters#Transporters/ other#Carbohydrate transporter	14	3611	16	8374	0.000326	0.006301	3.21
Metabolism#Amino acid/ nitrogen metabolism# Utilization of nitrogen sources other than amino	9	3611	9	8374	0.000513	0.009082	NA
Metabolism#Carbon metabolism#Utilization of specific carbon sources#Utilization of inositol	15	3611	18	8374	0.000568	0.009997	2.73
Cellular processes#Homeostasis#Metal ion homeostasis (K, Na, Ca, Mg)#Sodium uptake/ export	11	3611	12	8374	0.00069	0.011669	3.86
Metabolism#Carbon metabolism#Utilization of specific carbon sources#Utilization of arabiNA/ arabinose/ arabitol	11	3611	12	8374	0.00069	0.011669	3.86
Metabolism#Carbon metabolism#Utilization of specific carbon sources#Utilization of starch/ maltodextrin	11	3611	12	8374	0.00069	0.011669	3.86
Metabolism#Additional metabolic pathways# Phosphate metabolism	14	3611	17	8374	0.001111	0.016213	2.63
Cellular processes#Cell envelope and cell division# Cell wall synthesis#Biosynthesis of teichuronic acid	8	3611	8	8374	0.00119	0.016803	NA
Metabolism#Additional metabolic pathways# Biosynthesis of cell wall components# Biosynthesis of teichuronic acid	8	3611	8	8374	0.00119	0.016803	NA
Metabolism#Amino acid/ nitrogen metabolism# Utilization of amino acids#Utilization of branched-chain amino acids	17	3611	22	8374	0.00119	0.016803	2.17
Metabolism#Amino acid/ nitrogen metabolism# Utilization of nitrogen sources other than amino acids#	15	3611	19	8374	0.001619	0.021665	2.31
Metabolism#Nucleotide metabolism#Utilization of nucleotides	16	3611	21	8374	0.002167	0.02769	2.08
Cellular processes#Transporters# Transporters/ other#Other transporters	27	3611	41	8374	0.002736	0.032865	1.35

Appendix 10. – continued.

Information processing#Protein synthesis, modification and degradation#Proteolysis# Extracellular feeding proteases	7	3611	7	8374	0.002763	0.032865	NA
Information processing#Regulation of gene expression#Transcription factors and their control#Control of PRD-type regulators	7	3611	7	8374	0.002763	0.032865	NA
Information processing#Regulation of gene expression#Trigger enzyme#Trigger enzymes of the PTS that control the activity of PRD-	7	3611	7	8374	0.002763	0.032865	NA
Metabolism#Carbon metabolism#Utilization of specific carbon sources#Utilization of galactan	7	3611	7	8374	0.002763	0.032865	NA

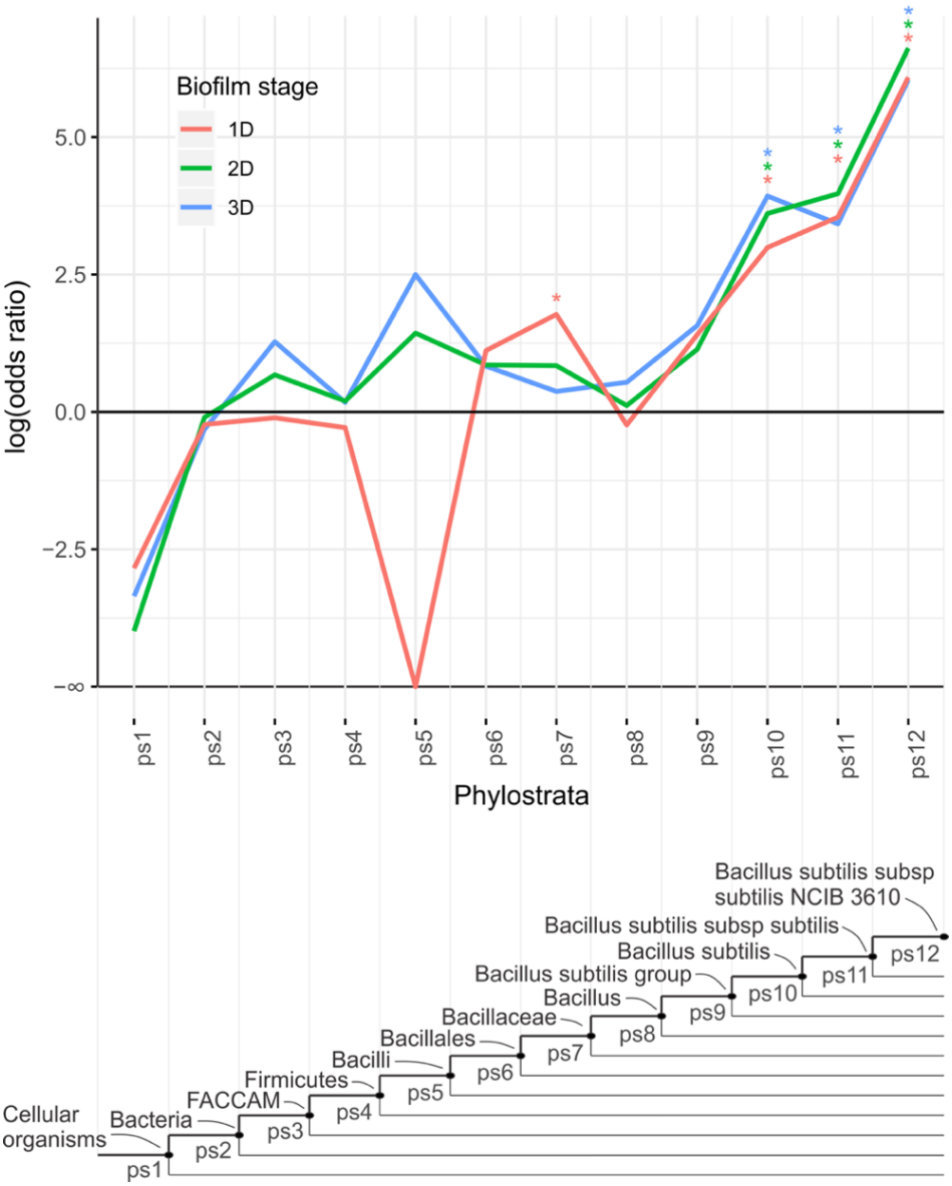
1M	q	s	h	t	p	padj	log_odds
Lifestyles#Coping with stress#General stress proteins (controlled by SigB)	78	783	159	8374	5.17E-39	2.21E-36	3.36
Lifestyles#Sporulation#Sporulation proteins# Small acid-soluble spore proteins	11	783	16	8374	1.26E-08	7.89E-07	4.43
Lifestyles#Coping with stress#Resistance against oxidative and electrophile stress	28	783	100	8374	6.82E-08	3.98E-06	1.95
Metabolism#Carbon metabolism#Utilization of specific carbon sources#Utilization of arabiNA/ arabinose/ arabitol	8	783	12	8374	1.99E-06	8.65E-05	4.29
Prophages and mobile genetic elements#Prophages# Skin element	17	783	60	8374	2.19E-05	0.000678	1.96
Metabolism#Carbon metabolism#Utilization of specific carbon sources#Utilization of ribose	5	783	6	8374	3.91E-05	0.001142	5.61
Lifestyles#Sporulation#Sporulation proteins# Spore coat protein/ based on similarity	4	783	5	8374	0.000351	0.006587	5.28
Metabolism#Carbon metabolism#Utilization of specific carbon sources#Utilization of glucarate/galactarate	4	783	6	8374	0.000976	0.014831	4.28
Cellular processes#Transporters# Phosphotransferase system# Sugar specific PTS proteins	8	783	24	8374	0.001064	0.015624	2.29
Metabolism#Carbon metabolism#Utilization of specific carbon sources#Utilization of pectin	7	783	19	8374	0.001122	0.016283	2.51

Appendix 10. – continued.

Lifestyles#Sporulation# Germination/ based on similarity	4	783	8	8374	0.003905	0.042694	3.28
Metabolism#Carbon metabolism#Utilization of specific carbon sources#Utilization of glucomanNA	4	783	8	8374	0.003905	0.042694	3.28

2M	**q**	**s**	**h**	**t**	**p**	**padj**	**log_odds**
Lifestyles#Sporulation#Sporulation proteins#Small acid-soluble spore proteins	12	148	16	8374	1.01E-18	1.45E-16	7.5
Prophages and mobile genetic elements#Mobile genetic elements#ICEBs1	8	148	25	8374	6.61E-09	4.71E-07	4.79
Lifestyles#Sporulation#Sporulation proteins#Newly identified sporulation proteins (based on transcription profiling)	16	148	173	8374	5.63E-08	3.36E-06	2.64
Lifestyles#Sporulation#Sporulation proteins# Spore coat protein/ based on similarity	4	148	5	8374	4.62E-07	2.28E-05	7.84
Prophages and mobile genetic elements# Prophages#SP-beta prophage	13	148	181	8374	1.75E-05	0.000549	2.21
Lifestyles#Sporulation#Sporulation proteins# Sporulation proteins/ other	16	148	274	8374	2.50E-05	0.000748	1.9
Metabolism#Carbon metabolism#Utilization of specific carbon sources#Utilization of fructose	3	148	8	8374	0.000284	0.005602	5.09
Cellular processes#Transporters# Phosphotransferase system#Sugar specific PTS proteins	4	148	24	8374	0.000756	0.012079	3.51
Lifestyles#Coping with stress#Acid stress proteins (controlled by YvrI-YvrHa)	2	148	5	8374	0.002996	0.034986	5.23
Prophages and mobile genetic elements# Prophages#Skin element	5	148	60	8374	0.004031	0.043693	2.38

	ps1	ps2	ps3	ps4	ps5	ps6	ps7	ps8	ps9	ps10	ps11	ps12
1D	10	7	1	5	0	2	6	2	3	3	9	10
2D	6	9	2	8	1	2	4	3	3	5	13	13
3D	9	8	3	8	2	2	3	4	4	6	10	11



Appendix 11. Distribution of functionally unannotated genes on the phylostratigraphic map. Evolutionary origin of genes that contribute to the enrichment of the functional term "No annotation" at 1D, 2D and 3D timepoints (see Figure 5). The table shows the number of genes without functional annotation per phylostratum for 1D, 2D and 3D timepoints. Phylostratigraphic enrichment is tested by one-tailed hypergeometric test and *p* values are adjusted for multiple testing (* $p < 0.05$). The abbreviation FACCAM (ps3) stands for Firmicutes, Actinobacteria, Chloroflexi, Cyanobacteria, Armatimonadates and Melainabacteria. Figure obtained from Futo *et al.* (2021).

8. CURRICULUM VITAE

Sara Koska obtained her MS in Animal Genetics and Breeding from the Faculty of Agriculture, University of Zagreb, in 2015, and MS in Experimental biology – zoology from the Faculty of Science, University of Zagreb, in 2016. In April 2017, she started working as a trainee in Hyla Association, where she participated in the field work and explored Croatian entomofauna and herpetofauna. In 2018, she enrolled in a doctoral study of biology at the Faculty of Science, University of Zagreb, and joined the Domazet-Lošo group at the Ruđer Bošković Institute, Division of Molecular Biology, Laboratory of Evolutionary Genetics. Working in the Domazet-Lošo group expanded her knowledge in the field of bioinformatics, statistics, as well as phylogenetics and evolutionary and developmental biology and genetics.