Philipps-Universität Marburg
Fachbereich Biologie

The Role of Genomic Context in
Bacterial Growth Homeostasis

Dissertation
zur
Erlangung des Doktorgrades
der Naturwissenschaften
(Dr. rer. nat.)

vorgelegt von

Andre Sim

aus Wellington, New
Zealand

Marburg, November
2019

Die Untersuchungen zur vorliegenden Arbeit wurden von August 2016 bis Oktober 2019 am LOEWE-Zentrum für Synthetische Mikrobiologie (SYNMIKRO) der Philipps-Universität Marburg unter der Leitung von Dr. Georg Fritz durchgeführt.

Erstgutachter: Prof. Dr. Torsten Waldminghaus

Zweitgutachter: Prof. Dr. Erhard Bremer

Vom Fachbereich Biologie angenommen am: 02.12.2019

Tag der mündlichen Prüfung: 11.12.2019

Eidesstattliche Erklärung

Hiermit erkläre ich, dass die vorliegende Dissertation:
"The Role of Genomic Context in Bacterial Growth Homeostasis" von mir selbstständig und ohne unerlaubte Hilfsmittel angefertigt wurde. Es wurden keine anderen als die von mir angegebenen Quellen verwendet. Zudem versichere ich, dass die Dissertation in dieser oder ähnlicher Form noch bei keiner anderen Hochschule eingereicht wurde.

……………………………………………………
Andre Sim, Marburg, 30 Oktober 2019

Table of Contents

Acknowledgements

I would like to begin these acknowledgements first with the people, that without their assistance, guidance and help, the construction of this thesis would not have been possible. Firstly, my supervisor Dr. Georg Fritz, thank you for allowing me to work on this project, in your lab and providing me scientific guidance and feedback. I would like to thank the students whom I supervised, Thomas Gotwig for his work on GenCoDB, Annis Newmann for her help in construction of the P_{uppS} reporter constructs and Jessica Bzdok, for her work on the *rasP* knock out mutants and microscopy. Of course, I need to thank everyone from my lab, past and present, who provided insightful discussions, scientific help and morale boosts when needed.

Next I would like to thank the team at Gießen, whose support allowed us to get GenCoDB online, Professor Alexander Goesmann, Burkhard Linke and Lukas Jelonek. Additionally, without the help of Dr. Hannes Link, Dr. Timo Glatter, Stefano Donati, and Anna Hakobyan the metabolomic and proteomic experiments would not have been possible. I would like to thank my thesis advisory committee who provided me critical feedback and the motivation to pursue new and exciting research questions, Professor Torsten Waldminghaus, Professor Erhard Bremer and Professor Gert Bange.

Finally, I would like to thank those that directly supported me in writing this thesis, namely Dr. Georg Fritz, Angelika Diehl, Lukas Hunziker and David Harvey for your corrections and critical reading of my thesis manuscript. Lastly, I would like to thank David Harvey and Zico, for important support throughout the course of my PhD. Without anyone on this list the outcome of this thesis would have been a lot grimmer, therefore, I thank you all one last time.

Zusammenfassung

Das Wachstum von Bakterien ist ein komplexes, aber gut organisiertes Spiel, bei dem ausreichende Mengen verschiedener Zellkomponenten produziert werden müssen, um Zellteilungen durchzuführen und den Zyklus zu wiederholen. Dabei kann vieles schief gehen - deshalb haben sich bei Bakterien mehrere Strategien entwickelt, um sicherzustellen, dass alle Prozesse synchronisiert ablaufen. Diese Problematik wird zusätzlich erschwert, da die Zellen ihre Wachstumsrate an ihre Lebensbedingungen anpassen, was wiederum die gesamte Zellphysiologie beeinflusst. Eine bemerkenswerte Änderung ist, dass mit zunehmender Nährstoffverfügbarkeit und -qualität die durchschnittliche Größe der Zellen und die Konzentration der Ribosomen in der Zelle steigt; letztere ermöglicht sowohl die Produktion der größten Makromolekülfraktion in der Zelle (Proteine) als auch mehr Ribosomen, die für ein schnelles Wachstum erforderlich ist. Mit der Zunahme des Volumens der Zelle kommt eine erforderliche Vergrößerung der Oberfläche, da ein Ungleichgewicht zwischen diesen beiden zu einem unhaltbaren Innendruck führen würde. Wie stellen Bakterien dann sicher, dass das Volumenwachstum mit der Produktion von Zellhüllenkomponenten synchronisiert wird, so dass die Zellhomöostase erhalten bleibt, insbesondere bei schwankender Wachstumsrate? Genomischer Kontext ist bekannt dafür die Koregulation von Genen zu unterstützen und dadurch ihre Expression auf verschiedene zelluläre Reize zu synchronisieren. Da das bakterielle Genom sehr unbeständig ist, deutet die Existenz konservierter genomischer Kontexte auf wichtige Ansatzpunkte der Koregulation hin. Könnte es sein, dass in diesen Genclustern das fehlende Puzzlestück zur Erklärung der Synchronisierung von Volumenwachstum und Oberflächenexpansion liegt?

Um diese Frage zu beantworten, werden in dieser Arbeit drei Fragestellungen bearbeitet. Zunächst entwickeln wir ein Genomvergleichstool (www.GenCoDB.org), dass die ständig wachsende Verfügbarkeit von sequenzierten bakteriellen Genomen nutzt, um die Analyse, den Vergleich und die Quantifizierung von Genomkontexten zu erleichtern. Dies beruht auf neuartigen Strategien, um die Breite der Genomdaten, die auf rechnerisch effiziente Weise verfügbar sind, zu berücksichtigen, die Wirkung von Probenahmeverzerrungen, die sich in den meisten bakteriellen Datensätzen finden lassen, zu verringern und sicherzustellen, dass Kandidaten für ihren evolutionären Kontext als signifikant angesehen werden können. Die Verfügbarkeit von GenCoDB wird die genomische Kontextforschung in der Mikrobiologiegemeinschaft erleichtern und den Zugang von Nicht-Bioinformatikern zu dieser Quelle wichtiger biologischer Daten verbessern.

Mit unseren neuen Erkenntnissen zu genomischen Nachbarschaftsdaten untersuchen wir anschließend die Evolution konservierter Gencluster und versuchen mögliche Kandidaten zur Regulation der Volumen-Oberfläche zu identifizieren. Beim Nachvollziehen der Verwandschaftsverhältnisse von Genclustern im gesamten Bakterienreich stellen wir fest, dass die Co-Orientierung stark konserviert ist, was jedoch weder den späteren Kontext um das Cluster herum noch die Expansion des Clusters beeinflusst. Wir finden heraus, dass die vertikale Übertragung und nicht der horizontale Gentransfer der treibende Faktor für das Auftreten von Genclustern in Chromosomen ist und dass Cluster an Origin und Terminator mit größerer Wahrscheinlichkeit instand gehalten werden. Schließlich stellen wir fest, dass trotz der scheinbaren Häufigkeit der Operon-Organisation in Genclustern, diese eher aufgrund anderer selektiver Belastungen wie Protein-Protein-Interaktionen innerhalb des Clusters und des essentiellen Statuses ihrer Gene aufrechterhalten werden, und dass Operons ein Produkt der Co-Lokalisierung über die evolutionäre Zeit zu sein scheinen.

Wir identifizieren einen einzelnen Gencluster-Kandidaten, der allen Anforderungen gerrecht wird, die unserer Meinung nach für die Homöostase des Zellwachstums der Oberflächenexpansion erforderlich sind. Die Anforderungen sind eine hohe Verbreitung innerhalb von Bakterien sowie eine Verbindung zwischen ribosomenassoziierten Proteinen (Wachstum) und Zellhüllesynthese. In Übereinstimmung mit unseren Evolutionsstudien finden wir heraus, dass das Cluster zwar ko-reguliert ist, dies aber nicht der selektive Druck zu sein scheint, der diese verschiedenen Prozesse zusammenführte. Stattdessen finden wir eine potenzielle Rolle des genomischen Channellings, das die Produktion von Pyrimidinen mit der Synthese der Zellhülle verknüpft, die von der Co-Lokalisierung dieses Clusters abhängig ist.

Insgesamt wird diese Arbeit das Verständnis der Chromosomenentwicklung in Bakterien und die potenziellen Auswirkungen des genomischen Kontextes auf die Metabolitenverwertung erweitern. Es stellt die Rolle von Operons und horizontalem Gentransfer bei der langfristigen Entwicklung der Genordnung in Frage und bietet eine neue quantitative und statistische Ressource, die den Zugang zu über 1,9 Millionen Gen-Nachbarschaften ermöglicht.

Abstract

The growth of bacteria is a complex but well-orchestrated dance involving the repetitive and reproducible production of their diverse cellular components in order to divide. A lot can go astray and therefore the cell has developed several strategies in order to ensure everything remains synchronized. This problem is only further complicated as the cells adjust their growth rate to their living conditions resulting in ripple effects throughout the cell physiology. One notable change is that as nutrient availability and quality increases so too does the average size and the concentration of ribosomes in the cell. The latter enables the production of the largest macromolecule faction in the cell (proteins) including the production of more ribosomes required to maintain the protein synthesis requirements. With the increase in volume of the cell comes a required increase in surface area, and a disbalance between these two would result in untenable levels of internal pressure. How then do bacteria ensure that volume growth is synchronized with the production of cell envelope components so that cell homeostasis is maintained, especially in the face of fluctuating growth rate? Genomic context is known to assist in co-regulation of genes thereby synchronizing them to respond to different cellular stimuli. As the bacterial genome is highly fluid, the existence of conserved genomic contexts suggests important loci of co-regulation. Could it be in these gene clusters that a possible link between growth and surface expansion is found?

To answer this question this thesis undertook three missions, firstly we established a genome comparison tool (www.GenCoDB.org) that will take advantage of the ever-growing availability of bacterial genomes to assist us in the analysis, comparison, and quantification of genome contexts. This will rely on novel strategies in order to: accommodate the breadth of genome data available in a computationally efficient manner, reduce the effect of sampling bias that plague most bacterial datasets and ensure candidates are considered significant for their evolutionary context. The availability of GenCoDB is sure to facilitate genomic context research in the microbiology community and improve accessibility to non-bioinformatics to this wellspring of important biological data.

With the swath of genomic neighbourhood data, we then sought to understand and analyse the evolution of conserved gene clusters in order to narrow down possible volume-surface regulating candidates. By tracking the evolution of gene clusters throughout the Bacteria kingdom we found that co-orientation is strongly conserved, however, this does not influence the subsequent context around the cluster nor the expansion of the cluster. We found that vertical transmission and not horizontal gene transfer was found to be the driving factor of gene cluster occurrence in chromosomes and that the origin and terminus are hotspots for cluster maintenance. Finally, we found that despite the apparent frequency of operon organization in gene clusters, gene clusters appear to be maintained due to other selective pressures such as within-cluster protein-protein interactions and the essential status of their genes. We suggest that operons are a consequence and not a cause co-localization over evolutionary time.

We identified a single gene cluster candidate that met all the requirements we believe are required for cell growth homeostasis of synchronized surface and volume expansion. These requirements were a broad conservation within Bacteria, and a connection between ribosome-associated proteins (growth) with cell envelope synthesizes. In agreement with our evolution studies we found that whilst the cluster was co-regulated this did not appear to be the selective pressure that brought these different processes together. Instead we found a potential role of genomic channelling, linking the production of pyrimidines with the synthesis of the cell envelope which is reliant on the co-localization of this cluster.

Together, this work will forward the understanding of chromosome evolution in Bacteria and the potential implications of genomic context in metabolite utilization. It challenges the roles that operons and horizontal gene transfer play in the long-term evolution of gene order and it provides a new quantitative and statistical resource providing access to over 1.9 million gene neighbourhoods.

1. Introduction

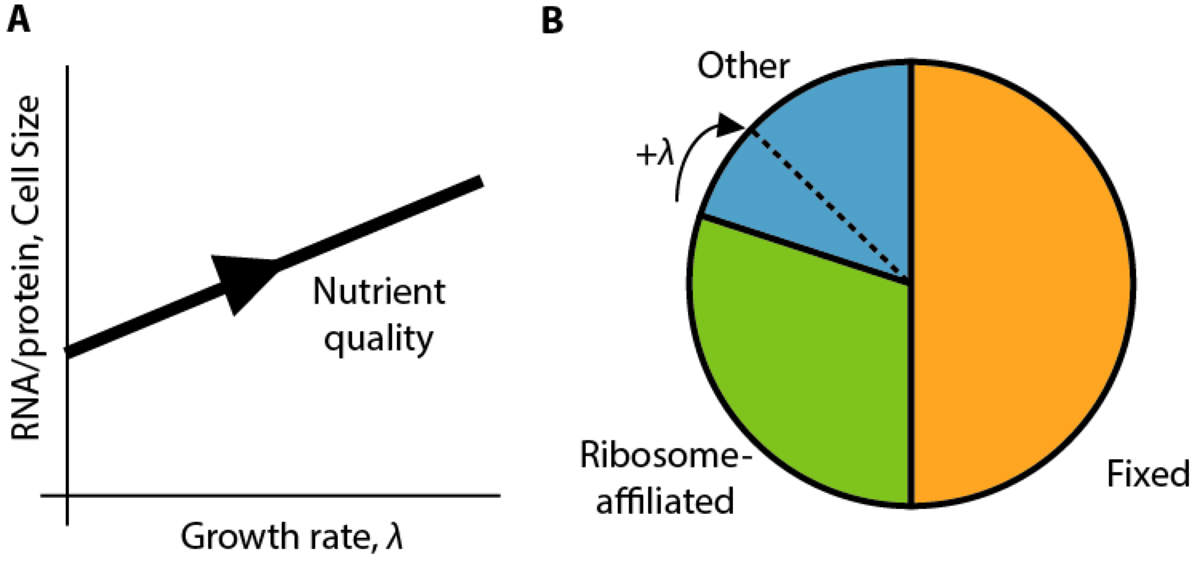
With the following introduction, we will introduce the concept of growth in bacteria. Most importantly we will focus on how during steady state growth, in a myriad of conditions, the cell ensures that the duplication of its cellular components such as proteins, DNA and cell envelope occur in step with the division of the cell. We will introduce the mechanisms cells have adapted in order to ensure these requirements are met even at high growth rates and in the face of perturbations. Then we will focus on how genomic context is a vehicle in which stoichiometry between different processes in the cell can be ensured as well as the different evolutionary pressures facing genome organization.

1.1 Cell growth in the context of bacteria

"What defines life?" - a hotly debated question amongst biologists, but one aspect in which they all agree upon is that without growth and self-replication, life could not persist (Koshland 2002). In the context of biology, growth can be defined both as the accumulation of cell mass, occurring when anabolism is greater than the rate of catabolism, and the proliferation of cells through division. These two processes are often tightly intertwined with each other as we will later detail in this chapter. Bacteria are the canonical model for understanding growth due to their ease of manipulation, their relative simplicity and rapid growth. Understanding the complexities of growth in bacteria is of great benefit to our society. Bacteria are being increasingly used as cell factories to produce metabolic products for medicinal, industrial and economical purposes (Kleerebezemab, Hols, and Hugenholtz 2000). Additionally, there is a strong need for research and development into bacteriostatic antibiotics, antibiotics that halt the growth of bacteria stopping their proliferation. Shortly after the discovery of culturable bacteria, their growth behaviour was classified into four distinct phases: lag, log/exponential, stationery and death phase. Only during the log phase would the cells be classified as growing and therefore this thesis will focus on this phase. In the log phase bacteria are both rapidly dividing and accumulating resources, increasing the total cell mass, as the name of the phase would suggest, exponentially. The frequency that a cell culture can double their mass is referred to as the growth rate, and this rate is determined by the growth conditions. Classically temperature, nutrient quality of the media, and environmental factors such as salinity, acidity and the presence of antimicrobial compounds can all modulate the growth rate of a cell. Typically the growth rate in natural environments would be governed by the limitation or quality of a nutrient, typically carbon, however, nitrogen, phosphorus and oxygen may also be limiting factors (Aldén, Demoling, and Bååth 2001). Changing solely the carbon source in the minimal medium M9 from a preferred sugar, such as glucose, to succinate can result in a doubling time increase from 70 to 134 minutes in *Escherichia coli* (Chang et al. 1999). Another important factor is the genetic context of the cell: different species even encountering the same growth conditions can have wildly different growth rates, for example, in a nutrient rich lysogeny broth (LB) *E. coli* doubles every 20 minutes whereas *Sinorhizobium meliloti* in the same media doubles every 140 minutes (Dai et al. 2018). One of the fastest-growing known bacteria, *Vibrio natriegens,* can reach a doubling time of 9.4 minutes in optimized media (Hoffart et al. 2017).

Here we should separate the concepts of cell division and cell growth rate. The growth rate is normally measured via the optical density of a cell culture which is a reasonable estimation of the number of cell mass doublings occurring in an hour. Simultaneously, bacteria undergo binary fission which involves the replication and segregation of their DNA into the production of two roughly equally sized daughter cells. When a cell divides and splits into two daughter cells, the number of cells in the population has doubled, however, the cell mass has remained identical. Having said that, a bacterial population must make sure that division and growth rate are synchronized. If division does not synchronize with mass doublings, for example division occurs consistently at a faster or slower rate, this would result in the progression towards infinitely small or large cells, as each generation would be a different size from the last. The processes in the cell which ensures this does not happen is referred to as "cell size homeostasis" (Taheri-Araghi et al. 2015). This highlights one of the many facets of the cell that needs to be adjusted with a change in growth rate.

Cell mass within bacteria is not homogenous and consists of several different cellular macromolecules which need to be duplicated before the coming division event, including DNA, RNA, proteins, and the cell envelope, including the cell wall and the phospholipid membrane(s). Each has their own mechanism of biosynthesis and herein lies the first challenge for the cell - as it is not simply a matter of globally increasing production of these macromolecules due to the many interdependencies in their synthesis. It was evidenced early on in microbial research that cellular composition (RNA, DNA, protein and the cell mass itself) is directly proportional to the growth rate and independent of the growth medium composition (Schaechter, Maaloe, and Kjeldgaard 1958). These observations led to the formation of mathematical laws which govern the relationship between the growth rate and the chemical composition of the cell. (Schaechter, Maaloe, and Kjeldgaard 1958). Described by the law is an exponential relationship between the growth rate and the respective concentrations of cellular components, and the exponent of these relationships varies with each component (Figure 1.1). RNA and ribosome concentrations increase with the growth rate, DNA concentration decreases, and protein concentration remains constant (Bremer and Dennis 2008). Strikingly, this does not happen when modulating the growth rate with temperature (Bremer and Dennis 2008). The question is why does each component behave differently? During rapid growth in *E. coli* half of the cell's biomass is protein (Bremer and Dennis 2008) of which the production consumes 75% of the cell's ATP budget (Russell and Cook 1995). Therefore, making protein synthesis very important in achieving fast growth rates. The protein content can be further divided into three fractions, ribosome-associated proteins, growth invariant proteins, and others (including constitutively expressed proteins) (Figure 1.1.B). In *E. coli* K12 MG1655 the fixed fraction is half of the protein fraction, whereas the non-fixed fraction, including constitutively expressed proteins inversely correlate with the growth rate (Figure 1.1.B) (Scott and Hwa 2011). In order to meet the protein synthesis demand, cells need ribosomes which consist of a mixture of RNA and ribosomal proteins, in other words the rate of protein synthesis is given by the concentration of available translating ribosomes. As a large portion of the protein mass is ribosome-associated proteins, this means there is a significant fraction of ribosomes involved in the proliferation of further ribosomal protein synthesis. Therefore, as growth rates increase, the fraction of proteins responsible for synthesizing further ribosomes needs to increase in order to accommodate the required protein synthesis load (Figure 1.1.B). This means that ribosomes synthesizing ribosomes is a key determinant of growth and the larger fraction a cell can devote towards ribosome self-reproduction the faster the cell can grow as a whole. Experiments in *E.coli* revealed that 85% of total cell RNA in *E. coli* is ribosomal RNA (rRNA) independent of growth rate (Bremer and Dennis 2008). Therefore, as the RNA/total protein ratio and the ribosomal protein/total protein ratio has a positive linear relationship with growth rate, this relationship mainly reflects the positive linear relationship between ribosomal mass and the growth rate (Figure 1.1.A) (Schaechter, Maaloe, and Kjeldgaard 1958). To ensure that the rRNA and ribosomal protein levels remain at stoichiometric proportions, ribosomal proteins are able to bind to the UTR of their own mRNA (Nomura, Gourse, and Baughman 1984). Typically they bind rRNA with a high affinity, however in cases of insufficient rRNA to bind, they bind their own mRNA and thereby autoregulate their synthesis until sufficient rRNA can be produced (Nomura, Gourse, and Baughman 1984). This ensures stoichiometric levels of the two ribosome components irrespective of the growth rate.



1 - Figure 1.1 - Growth laws in bacteria

A. As growth rate (*λ*) increases due to higher nutrient quality, so to does the RNA/protein and cell size, which both correlate linearly. This is highlighted in the increase of RNA (mainly ribosomes) that the cell requires in order to meet protein synthesis demands. **B.** The cellular protein content can be abstracted into three fractions, a fixed fraction which is invariable to changes in growth rate (orange), the ribosome-affiliated proteins which increase with growth rate (green) and other proteins inversely correlated with growth rate (blue). The dotted line represents how the ribosome-affiliated fraction consumes the "other" fraction as growth rate increases. Adapted from (Scott and Hwa 2011)

Further complexity was found in the same work by Schaechter et al. who also observed another fundamental bacterial growth law: cell volume and thereby size is linearly correlated with the growth rate (Figure 1.1.A). As the growth rate of a cell increases so does the cell volume, again independent of the composition of the growth media. This has many implications to the regulation and concentration of macromolecules in the cell as the volume increases. However, it was found that independent of the size of the cell, the protein concentration continued to remain constant (X.-Y. Zheng and O'Shea 2017), correlating the expansion of volume with the active ribosome pool. The question of why bacteria increase their size with an increased growth rate/nutrient availability is still unanswered. The first of two potential explanations posit that it is to accommodate the volume of DNA, which at high growth rates undergoes multifork replication and therefore requires significantly more space. This hypothesis is supported by the fact that bacteria not performing multifork replication, such as *C.crescentus,* do not alter their cell size with their growth rate (Campos et al. 2014). An alternative explanation suggests that increasing cell size during high growth conditions is a crude method of storing nutrients in order to survive and adjust to future periods of starvation (Westfall and Levin 2017). From these observations it has become clear that in the variation of growth rate the relative concentrations of different cellular macromolecules are finely tuned to maintain such mass doubling speeds. At specific growth rates the number of ribosome complexes needs to reflect the number of active RNA polymerases and other complexes in tight stoichiometry. However, it is not only macromolecules that must be synchronized, but the sub-processes which synthesize them as well, namely translation, transcription, and DNA replication. With the association between growth rate and cell size this goes further still, requiring that there must be stoichiometry and synchronization between even more encapsulating processes such as cell volume and cell surface growth. How then can the bacteria synchronize everything at once whilst keep pace with the ticking clock that is division in the face of environmental changes, stresses and perturbants?

1.2 Synchronization strategies in bacteria

The requirements for cell division can be loosely divided into three sections: replication of the chromosome(s), production of twice the number of cellular components and the act of division itself. With the complexity of each of these three tasks, this begs the question, what strategies do bacteria use in order to synchronize these processes? As we have already discussed at length how the different macromolecules are regulated with growth rate, we will next discuss DNA replication. In the context of cell division the replication is divided into three periods, the B period, spanning the birth of the cell and division, the C period, which is the time required to completely duplicate the DNA and the D period; the time between finishing DNA replication and division (Bremer and Dennis 2008). In *E. coli* the time it takes to replicate the genome is approximately 41 minutes, however this does vary based on growth rate slows down in poor growth conditions (Cooper and Helmstetter 1968). We have already mentioned that *E. coli* can achieve doubling times of 20 minutes, however if the doubling time of a cell is less than the C period, how could the cell possibly produce enough DNA in order to divide in time? To solve this issue and as alluded to earlier, many bacteria undergo what is referred to as multifork replication, where the replication of the DNA is initiated before the previous round has finished. In *Pectobacterium carotovorum this* can result in up to 30 replisomes on the DNA at one time (Couturier and Rocha 2006). Through multifork replication, bacteria can ensure that at division, at least one round of replication is finished once cell mass has doubled division is required (Donachie 1968). A consequence of multifork replication is that there an increase in gene dosage of genes close to the origin of replication (ori) (Soler-Bistué, Timmermans, and Mazel 2017). Bacteria have taken advantage of this by biasing the which genes are localized near the origin. It has been found that several translation- or ribosome-associated genes are located there in many bacteria (Soler-Bistué, Timmermans, and Mazel 2017). This provides our first clue to the relevance of genome organization in the synchronization of essential pathways during cell growth, which we will discuss in more depth later. Indeed, the amount of protein found in a cell has been found to be a function of the number of replication origins, thereby extrapolating from this, the amount of DNA in the cell is correlated to the amount of protein in the cell (Donachie 1968). What then correlates the growth rate with the onset of multifork replication? Although The precise mechanisms of initiation remain elusive but we know that replication initiation occurs at relatively constant cell volumes respective to the current number of origins of replication in the cell(Donachie 1968; Hill et al. 2012; Wold et al. 1994). This initiation volume is independent of both the growth rate or the birth size of the cell (Wallden et al. 2016). One hypothesis suggests that this is because the expression of proteins involved in the initiation of replication, an example from *E.coli* being the replication initiation protein DnaA, are autoregulated maintaining their concentration independent of the volume or growth rate of the cell (Skarstad and Katayama 2013). Initiation would then occur through the accumulation of a fixed critical amount of replication initiator at the chromosome origin. Then with initiation, the proteins are diluted between the now multiple origins and must accumulate again before a new replication fork is initiated (Si et al. 2017). When there are already multiple origins, they are required to initiate relatively simultaneously as they begin to become hemimethylated and protected from further re-initiation (M. Lu et al. 1994).

After finalization of macromolecule synthesis, the cell needs to undergo division, dividing these components between the daughter cells. As we explained earlier in the introduction, division needs to be coordinated with the growth rate or else the size of the cells would inflate or shrink during each generation. There has been much research as to how the cell decides when to trigger division, and it was thought to occur through one of three different models: the adder, sizer and timer models (Taheri-Araghi et al. 2015). Under the adder model, cells divide after adding a fixed amount of size to their initial size, under sizer, the cell waits until they are a predetermined size before initiating division and with the timer model the cell initiates a fixed amount of time after birth (Taheri-Araghi et al. 2015). It was shown that under the majority of growth conditions bacteria divide under adder model (Amir 2014), however at slow growth rates this breaks down and division begins as one would expect in the sizer model (Wallden et al. 2016). The adder model can ensure cell size homeostasis despite the stochasticity of cell division, as initially smaller cells add proportionally larger amounts of cell mass before dividing and vice versa with initially larger cells, resulting in them converging in a similar growth-rate-defined size (Lin and Amir 2017). Exactly how the growth rate determines what the fixed size should be mechanistically is currently not known. Along with the increase in volume and size comes an increase in surface size. This must be matched by the production of the cell envelope as disbalance between volume and surface may lead to unsustainable levels of internal pressure in the cell or an unstable wall or membrane (Koch 1985). It is seen that disrupting the balance between cell wall synthesis and cell volume growth by cell wall targeting antibiotics has a greater effect at faster growth rates (Aldridge et al. 2012). As growth rate is coupled with: cell size, the total levels of ribosomes and the protein fraction in the cell, this suggests a cellular mechanism which synchronizes the volume expansion of the cell, and subsequently the production of ribosomes, with the production of the cell envelope. To this end we will explore different expression strategies bacteria use in order to maintain cell homeostasis.

Having mechanisms in place to maintain stoichiometry during different growth rates is important, however these need to persist not just in perfect growth conditions but also in natural contexts. Contexts where the growth homeostasis relationship outlined by the bacterial growth laws can be and are challenged. There are many different strategies and mechanisms utilized by bacteria to counteract different perturbants such as salt, iron, pH, and antibiotics to name a few examples. These are often regulated by signal transduction modules where the perturbant or stimuli is detected by one module and conveys the signal to an intracellular responder that can elicit changes in the transcriptional profile of the cell. An example of this is the stress response to perturbations of cell envelope synthesis in *B. subtilis* which consists of four sigma factors σ^W, σ^X, σ^M, σ^V each with overlapping but different stimuli. For briefness we will just cover the roles of σ^W and σ^M. σ^W is stimulated by membrane-active agents such as detergents and has a regulon consisting of 60-90 genes (Helmann 2006). The regulon consists of genes that provide resistance to antimicrobial agents, for example *fosB* which inactivates the MurA-targeting antibiotic fosfomycin (Cao et al.

2001). Furthermore genes involved in reshaping the membrane lipid composition are also activated. This leads to decreased membrane fluidity thereby providing long term resistance to membrane-active stressors allowing the cell to continue growing (Kingston, Subramanian, and Rock 2011). In contrast to σ^W, σ^M does not solely focus on upregulating genes involved in the detoxification and resistance to antimicrobial agents when peptidoglycan synthesis is inhibited (Eiamphungporn and Helmann 2008). Instead several genes within the peptidoglycan biosynthesis pathway such as *murG* and *amj* are upregulated to maintain the unperturbed synthesis rate and keep cell homeostasis (Eiamphungporn and Helmann 2008). The widespread use of sigma factors underpins the benefit of coordinating gene expression in order to synchronize the activity of many proteins.

Sigma factors are one of the ways bacteria coordinate gene expression and this is often required when responding to specific stimuli or changes in their environment. At the transcriptional level, transcription factors and sigma factors are utilized by the cell. Transcription factors are usually sequence-specific DNA-binding proteins and respond to stimuli to regulate transcription of a gene. Transcription factors may increase transcription by making the promoter region more accessible to the RNA polymerase, stabilizing its binding or recruiting other co-activators (Balleza et al. 2009). Conversely, they may also reduce transcription by blocking access of a promoter to RNA polymerase (Balleza et al. 2009). We discussed the role of some sigma factors in the previous paragraph; however their mode of function differs to transcription factors. Unlike transcription factors, sigma factors are required for transcription in bacteria. They associate themselves with the RNA polymerase complex and influence the promoter sequence affinity of RNA polymerase (Maeda, Fujita, and Ishihama 2000). They are often expressed constitutively and released from an anti-sigma factor upon recognition of the stimuli resulting in a rapid response. Genes that are collectively regulated by the same transcription factors or sigma factors are denoted as regulons and can be restrictive, affecting only two genes, or globally changing the expression of over 500 genes. Genes in a regulon can be dispersed across the genome. This strategy of co-regulation has been shown to be noisy, meaning that there is a lot of variability in the expression of genes responding to the same signal (de Lorenzo and Pérez-Martín 1996) This is often dependent on the individual genomic context of each gene, for example from gene dosage from multifork replication (Sauer et al. 2016), genome supercoiling (Dorman 2006) and upstream and downstream transcription events. The pioneering work of Monod (Jacob and Monod 1961) revealed the existence of operons in bacterial genomes. The canonical operon is usually described as a group of functionally similar genes (acting, e.g., in the same metabolic pathway), which are controlled by a single promoter, are close together, are all orientated in the same direction, terminate at a single transcription terminator and are transcribed at similar levels (Laing et al. 2006; Jacob and Monod 1961; Salgado et al. 2000). Operons reduce the noise in expression variability between the co-transcribed genes. This enables tight synchronization between the gene products (Ray and Igoshin 2012). As operons are also co-localized/clustered on the genome, the other aforementioned effects would not impose variability between expression of different genes. Operons are usually conserved in multiple species suggesting importance in the tight regulation they provide which can be of critical importance in the formation of multisubunit protein complexes. Therefore, regulation via operon organization makes for an interesting candidate in the synchronisation of surface and volume growth.

Operon transcription is not the only advantage of co-localizing genes on a genome affords to bacteria. It has been observed that several genes which encode protein-complexes and biosynthetic pathways are also found together in gene clusters (Dandekar et al. 1998; Fani, Brilli, and Liò 2005). The benefits of this are explained by the molarity model which posits that co-transcription and translation result in increased local concentrations of gene products (Gómez, Cases, and Valencia 2004). This thereby facilitates interaction or complex formation between the proteins as they are more likely to find their corresponding partner. In *Mycoplasma genitalium* it was shown that gene clustering was present in over a third of all known functional protein-protein interactions (Huynen et al. 2000). Therefore, it could be suggested that genomic context conservation is important to maintaining physically interacting proteins and therefore generates a strong selective force, especially on proteins involved in crucial physiological functions. The localization of proteins may be especially relevant in synchronizing enzyme activity levels, for example with moonlighting proteins. Moonlighting proteins are enzymes with more than one function (Huberts and van der Klei 2010). When moonlighting proteins interact with other proteins and processes this creates a potential synchronizing link between the two processes. An example is glucosyltransferase UgtP in *B. subtilis* (Weart et al. 2007). Firstly, OpgH is required for the synthesis for a gram-positive cell wall component, the diglycosyl-diacylglycerol anchor for lipoteichoic acid. Additionally however, when growing in nutrient rich medium and one of its substrates, UDP-glucose, is abundant, OpgH interacts with the cell division protein FtsZ, resulting in a delay of division and increasing cell size (Weart et al. 2007). It is therefore speculated that through OpgH, cell wall synthesis, carbon metabolism and cell division could be synchronized. As co-localised genes are often localised together (Mingorance, Tamames, and Vicente 2004) this suggests further moonlighting (and therefore) synchronizing strategies could be found at the genomic context level.

1.3 Bacterial genomic context evolution

As we are interested in the synchronization of highly conserved processes in bacteria (cell wall, synthesis, translation, etc), it would be expected that such a context would be found conserved across the Bacterial kingdom. Therefore, to understand the role synchrony plays in genomic context we must first understand the forces which shape and maintain bacterial chromosome organization. Gene order (synteny) is notably poorly conserved between bacterial species (Mushegian and Koonin 1996; Itoh et al. 1999; Dandekar et al. 1998) and that disruption of gene order occurred at a faster rate than the mutation of protein amino acid sequences (Wolf, Rogozin, Kondrashov, et al. 2001). However, despite the high rate of gene shuffling, bacterial genomes, even those distantly related to each other, do not appear to be independent collections of randomly ordered genes, and there are indeed several conserved genomic contexts and genene pairs (also referred to as gene neighbourhoods or gene clusters) that have been identified (Wolf, Rogozin, Kondrashov, et al. 2001). Examples of such cases include a mega-ribosome cluster consisting of over 20 ribosomal genes (Ohkubo et al. 1987), the genes encoding the ATP synthase complex (McCarn et al. 1988) and the division and cell wall (DCW) cluster (Pucci et al. 1997). The existence of these conserved clusters in defiance of the observed decline of context conservation at larger genomic scales indeed suggests that there must be fitness benefits conferred by these genomic arrangements such as cell process synchrony (J. Lawrence 1999; J. G. Lawrence and Roth 1996). The selective benefits of synchrony is only one plausible explanation for gene clustering. Further explanations can be broken down into three partially overlapping categories: the mechanisms which brought the genes together (Fani, Brilli, and Liò 2005; Touchon and Rocha 2016; Ream, Bankapur, and Friedberg 2015), which evolutionary forces are important in maintaining the cluster (Fang, Rocha, and Danchin 2008; Oliveira et al. 2017) and how these forces impact the biological function of the cell

(Tamames et al. 2001; Wells, Bergendahl, and Marsh 2016; Mingorance, Tamames, and Vicente 2004; Dandekar et al. 1998).

The fluidity of the bacterial genome stems from the many mechanisms which can alter its organization. There are both intrinsic factors resulting from the cell's own (error-prone) processes, and from external factors where DNA is taken up or forcefully inserted into the genome. Intrinsic rearrangements typically occur during the malfunction of regular cellular machinery leading to the inversion, deletion, duplication or translocation parts of the chromosome, referred to collectively as rearrangement events. The major source for such events is due to homologous recombination, which is a DNA repair process found in bacteria. Homologous recombination occurs after DNA damage has been detected involving either a double- or single-strand DNA break (Dillingham and Kowalczykowski 2008). DNA damage can be induced through several factors including UV light, radiation, restriction enzymes and chemical mutagens. RecBCD or RecF pathways, for double- or single-strand breaks respectively, are used to repair the gap (Smith 2012). Whilst both pathways utilize different proteins and mechanisms initially to unwind and degrade one end of DNA near the DNA break they both result in single stranded DNA 3’ end covered in RecA protein (Smith 2012). The RecA-coated nucleoprotein filament searches for homologous stretches of DNA and then undergoes strand invasion where it moves into the homologous recipient DNA duplex. This forms a D-loop which can be resolved in two ways. First the loop is cut resulting in swapping the strands between the two DNA molecules, and the gaps can be filled with DNA polymerase leading to two altered DNA fragments. Alternatively, the invading 3’ end can prime DNA synthesis and form a replication fork. Rearrangements occur when homology-dependant recombination-repair machinery recruits a similar but incorrect match as the repair template. The resolution of the mismatch can then lead to inversions, deletions or duplications depending on the orientation and location of the incorrect repair template and the damaged DNA (West 2003). Due to homologous recombination it has been shown that during stalls of the replication forks, inversions between the two sides of the genome are significantly more likely (Tillier and Collins 2000), resulting in high frequencies of inversions being centered on the origin and terminus of the chromosome (Suyama and Bork 2001). Due to high levels of sequence repetition, rearrangements occur especially frequently with repetitive genomic elements (Achaz et al. 2003). Direct repeats result in deletion of the sequence between them whilst inverted repeats lead to inversion. Another pseudo-intrinsic factor shaping chromosome organization are the presence of transposable elements (transposons). Whilst transposons were originally identified in plants (McCLINTOCK 1950), they have been found in nearly all organisms having evolutionarily important roles in genome construction. Transposons can be classified into different classes; however all are mobile genetic elements which often persist in the genome in a selfish way and can either be self-sufficient or require the presence of other transposable elements to move. They usually are surrounded by flanking regions which either facilitate: transcription, retrotranscription and reinsertion into the genome (class I); or excision and insertion (class II) (Kapitonov and Jurka 2008). In bacteria, transposons have been shown to frequently carry genes involved in other functions such as antibiotic resistance. This can also occur between chromosomal DNA and plasmids resulting in a mechanism in which foreign genes can be inserted into the genome.

The uptake of external genes can also occur through other mechanisms in bacteria, for example transformation, conjugation and transduction. Transformation is the uptake and integration of extracellular DNA performed by many bacteria. One example is found in *B. subtilis,* which enters a state of competence as they leave the exponential phase. While competent, they begin uptaking DNA (Solomon and Grossman 1996). DNA is then integrated based on homologous recombination and therefore usually is sourced from bacteria of the same species. There are infrequent exceptions where non-homologous DNA is integrated resulting in the insertion of foreign DNA. Conjugation involves the extended contact between two cells, a donor and recipient. A sex pilus is built between the two cells and DNA (usually an episome) is transferred from the host to recipient and can be incorporated into the host. The conjugation of Mycobacteria is chromosome- instead of plasmid-based (Derbyshire and Gray 2014). Transduction is the movement of genetic information between bacteria through a virus or viral vector. Firstly, the host cell DNA is packaged into viral capsids which are released through lysis. When the phage capsids infect a new host, the new DNA can be integrated into the host's DNA. Through these mechanisms multiple genes can be transferred at once across species boundaries, what is referred to as horizontal gene transfer (HGT). Through HGT gene clusters can be found over large taxonomic distances.

We have outlined both how fluid the bacterial chromosome is and the multiple mechanisms that create such genetic variability. However, without strong selective forces, any formed gene clusters will not be maintained long in a population. There are several hypotheses as to why gene clusters may arise despite strong chromosomal fluidity. One such model states that tight compaction of genes improves the rate for which HGT may occur between species. HGT subsequently results in the propagation of an apparent “conserved gene cluster” to a diverse array of taxa (J. G. Lawrence and Roth 1996). Under this model only non-essential genes are likely to cluster together and be transferred, unless the essentiality of the gene was developed after transfer to the recipient genome. Additionally, groups of genes that are laterally transferred are often similar in function as they must provide a selectable phenotype to the recipient – leaving little allowance for genes of divergent function to cluster together through this mechanism. The Fisher model works on the presumption that co-localized genes also co-adapted (for example corresponding amino acids in protein-protein interactions) and thereby, having them situated proximately to each other reduces the chance they will be separated by recombination events (Fisher 1929). This may then result in additional genes co-adapting and increasing the size of the inseparable cluster. In gene clusters which have already been discovered in bacteria, there was an enrichment in genes essential for cell growth. This has been explained for two reasons. By clustering essential genes, this reduces the probability that deleterious rearrangement events will span regions containing essential genes (Fang, Rocha, and Danchin 2008). Furthermore, as essential genes are less likely to participate in rearrangement events, particularly deletion events, this results in the deletable distances between essential genes shrinking over generations resulting in the auto-coalescence of multiple essential genes (Fang, Rocha, and Danchin 2008).

In the previous section we discussed the fitness benefits of operons and protein-protein interactions related to co-localization on the genome. Similar to the concept of operon transcription, co-expression of genes is another hypothezied evolutionary force. The “piece-wise” model explains the formation of more complex operons, which may contain a diverse array of gene functions (Fani, Brilli, and Liò 2005). It explains that small clusters, formed through other mechanisms, such as those mentioned above and below, may themselves cluster under a need for common regulation. This then leads to the build-up of larger and larger gene clusters. Coexpression is classed as a short term selective force but over long evolutionary time periods would not be able to maintain gene clusters together (Fondi, Emiliani, and Fani 2009). Another study found that coregulation and protein-protein interactions could explain ⅔ of gene pair clustering (Fang, Rocha, and Danchin 2008), this leaves however a third of gene pairs unexplained. These are only a selection of hypotheses regarding the formation and maintenance of gene clusters and it is possible we have not yet scratched the surface. One example was pointed out by Tamames et al (2011), who showed that organisms that had lost the DCW cluster, yet maintained all DCW genes independently, also lost their rod shape. If the localization of the cluster is required for the rod shape, or if selective advantage of the DCW cluster is only present in rod shaped bacteria, is currently an area of active research.

Despite the fluidity of the bacterial genome, several conserved neighbourhoods of genes have been found across the bacterial kingdom. Because of their evolutionary significance, conserved gene clusters have begun to be exploited for several purposes in biological research. For example Dandekar et al. showed that conserved gene-pair co-localization could be used to predict protein-protein interaction partners. Another example was the annotation of four GTPases as having a role in translation solely based on their conserved genome context (Wolf, Rogozin, Kondrashov, et al. 2001). Furthermore, this relationship has been exploited in order to predict the functional activities of unannotated proteins (R. Overbeek et al. 1999; Huynen et al. 2000). While these examples highlight the usefulness of genomic context conservation for various research questions, access to quantitative, statistical data on genomic context conservation is relatively limited – especially for scientists without a bioinformatics background. Currently there are numerous and excellent genomic databases for bacteria available, including microbe-wide datasets such as Microbes Online (Dehal et al. 2010), JGI (Grigoriev et al. 2012), NCBI (NCBI Resource Coordinators 2018), StringDB (Snel et al. 2000) and species-specific databases such as SubtiWiki for *B. subtilis* (Michna et al. 2016) or EcoCyc for *E. coli (Keseler et al. 2017)*. All these databases allow the visualization of the genomic context around a chosen gene in one form or another, affirming the ubiquitous need for this information. As examples, Microbes Online and JGI allow users to view multiple contexts concurrently (each genome is displayed as a separate line). StringDB takes another approach by displaying only genes which are frequently observed at each position on the neighbourhood for different taxonomic groupings. Through colour-coding of orthologous genes in the context of a gene of interest, these tools enable a semi-quantitative description of potential conservation patterns. This is done by manual counting of conserved genes across a limited number of genomes. However, this visual inspection of individual genomes is time-consuming and error-prone, typically preventing a statistical analysis of thousands of taxonomically diverse genomes, as required for rigorous conclusions about genomic context conservation. Another pitfall is that genomic databases are naturally biased towards species of high medical, biotechnological or academic interest, which can obscure statistical analyses of genetic context conservation. For example, if genomes within a subset are closely related, observed similarities between genomes are less likely due to appear because of selective pressures maintaining such an arrangement, but as a mere consequence of insufficient evolutionary time for genomic rearrangement events to have occurred. If not taken into account, this bias would result in false positive conserved neighbours identified in highly sequenced bacterial clades, such as Proteobacteria, and false negative neighbours not detected in less sequenced clades, such as Bacteroidetes. Accordingly, accounting for the non-random distribution of bacterial genomes is paramount for a meaningful analysis of genomic context conservation, but to date there is no publicly available database that provides such statistical analyses.

Despite the similarities in DNA code and transcription/translation apparatuses, the evolutionary forces that act upon eukaryotic and prokaryotic genomes is very different (Michalak 2008). Eukaryotes lack the ability (in most cases) to transcribe polycistronic transcripts and therefore do not have the operon-level organization found in bacteria genomes. Instead eukaryotes often have gene clusters which are also coregulated but comprise genes that have been duplicated and diverged. One such example is canonically represented by the β-globin gene cluster. The genes in this cluster are controlled by a local cis-acting sequence upstream of the cluster (Tanimoto et al. 1999). These are then often surrounded by chromatin insulators which lead to regions of gene silencing via heterochromatization (as opposed to the actively transcripted euchromatic state) (Gerasimova and Corces 2001). Chromatin is a term for packed eukaryotic nuclear DNA which is wrapped around a protein octamer referred to as a histone. The modifications of the histones control the state of the chromatin and particular modifications are associated with different expression patterns. For example methylation of Lys9, Lys27 and Lys35 of histone H3 is linked to heterochromatization (Lachner and Jenuwein 2002), whereas methylation of histone H3 at Lys4 is associated with euchromatization (Santos-Rosa et al. 2002; Zegerman et al. 2002). In addition to histone modifications, the DNA can also be methylated which in many organisms also results in heterochromatinization and gene silencing (Geiman and Robertson 2002). Both DNA methylation and histone recruitment can be modulated by DNA sequence specific factors, which often lead to the spread of this signal in both directions often covering large sections in similar regulation patterns. This therefore leads to genes clustering together based temporal and spatial expression requirements.

1.4 Project aims

The aim of this project is to further understand how, at diverse growth rates, bacteria maintain stoichiometry between volume and surface growth. Early observations of growing bacteria revealed that independent of the composition of the media, growth rate is modulated so it correlates the volume/size of the cell (Schaechter, Maaloe, and Kjeldgaard 1958). Measurements of the chemical composition of these growing cells highlighted a linear relationship between the growth rate and the number of active ribosomes, specifically the pool of ribosomes required to be synthesizing other ribosomes in order to match the protein translation demand of the growing cell. With this increase in volume, naturally comes an increase in the surface area of the cell and subsequently requirements for additional cell envelope components such as peptidoglycan and phospholipids. Whilst much work has looked into how the processes of DNA replication and division are regulated under the different growth rates, how cell envelope biosynthesis is regulated in relation to growth rate is not as well understood. Specifically, we will explore the role genomic context may have in connecting these two processes. To accomplish this, first we will develop a platform to allow us to quickly, quantitatively and statistically peruse the genomic contexts across the broad range of genetic diversity within Bacteria. The implementation of thousands of bacterial genomes will facilitate the analysis of genome context evolution and determine which gene neighbourhoods can be considered interesting candidates in the context of evolution. Our aim is to find gene neighbourhoods containing ribosome-associated genes co-localized with genes involved in the biosynthesis or homeostasis of the cell well, one that is well conserved across the Bacterial kingdom. Upon finding a candidate gene cluster we will explore its possible role in synchronizing surface and volume growth. Firstly, we will bioinformatically measure the correlation of expression between the genes of the gene cluster in multiple organisms where the cluster is conserved. We will test for the presence of noise-reducing operons in the gene cluster that may link ribosomal and cell envelope genes. Finally, we will disrupt the co-localization and observe how cell growth in different media is affected.

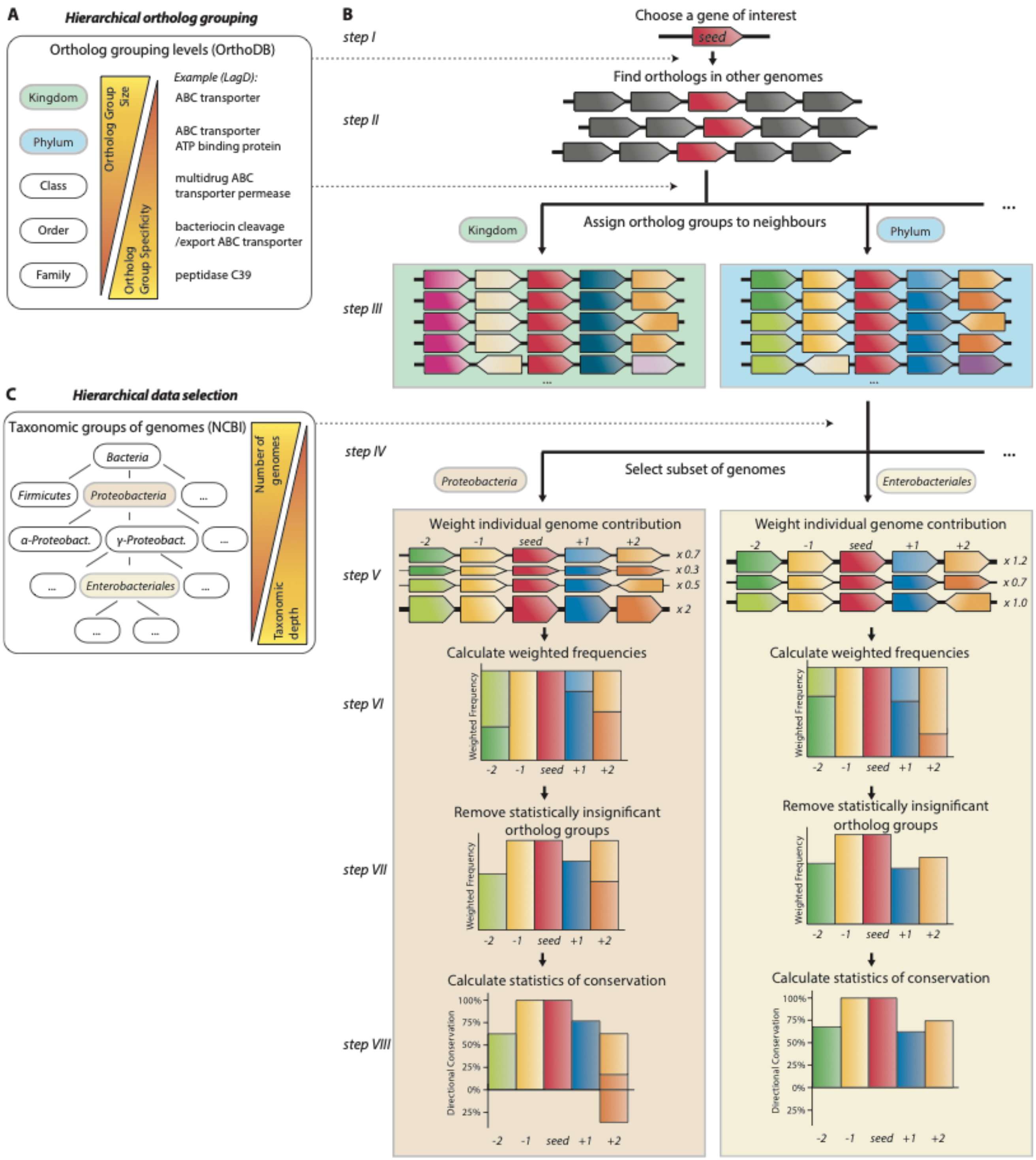
2. GenCoDB - A statistical tool for genetic context conservation analyses in bacterial genomes

In this chapter we will present GenCoDB (Genomic Context Database), an easy-access tool enabling the rigorous study of genetic context conservation in bacteria. GenCoDB implements statistical analyses that correct for sequencing bias and allows for an unprecedented resolution of genetic context conservation at different taxonomic levels and in individual clades. By exploiting the hierarchical ortholog group definitions from OrthoDB (29), GenCoDB categorizes genes into orthologous groups with user-adjustable levels of fine-graining, thereby permitting deep insights into the conservation of genes encoding either broad or more specialized biological functions. In GenCoDB the user can analyse the genomic context of a given gene via three complementary views, comprising an enhanced genome-by-genome view, a statistical neighbourhood view, as well as an evolutionary view showing the conserved genetic context along the bacterial tree of life. By combining the best features of conserved gene context visualization from Microbes Online, JGI and StringDB, as well as by adding new quantitative statistical analyses, GenCoDB fills a gap in the space of current databases without creating redundancies with previous tools. This database is publicly available at the Genomic Context Conservation Database (www.gencodb.org).

2.1 Data collection

Gene neighbourhood analysis

To compare gene neighbourhoods between different organisms, it is necessary to define which genes encode for similar proteins across the different genomes. These orthologs are traditionally identified by clustering genes based on similarity in the encoded protein sequence, as in the commonly used PFAM and COG ortholog group definitions (El-Gebali et al. 2019; Tatusov et al. 2000). However, the definition of whether two genes are orthologous is subjective to the research question and there are, in principle, many levels of course- or fine-graining that can be applied. This is especially apparent with highly abundant but diverse genes, encoding for instance ABC transporters, for which very small differences in protein structure can result in the import/export of vastly different substrates completely changing their cellular role. Thus, to allow for a differentiated definition of orthologous genes, GenCoDB is built upon the hierarchical ortholog grouping of OrthoDB (Kriventseva et al. 2019), in which each protein sequence was clustered multiple times with different subsets of protein sequences belonging to organisms of different taxonomic groups. Consequently, by being exposed to either more closely or more divergent sequences, every protein is assigned to different ortholog groups of different levels of course-graining – trading specificity for generality (Figure 2.1A). For instance, as shown in Figure 2.1A, the ATP-binding protein LagD is assigned a general ABC transporter group at the kingdom level, but a much more specific group, bacteriocin cleavage/export ABC transporter, at the order level. Notably the exposure of different subsets of sequences resulted in LagD being grouped with permeases (despite being an ATPase) at the class level, highlighting the need to modulate the sensitivity of ortholog grouping classifications.


A
Hierarchical ortholog grouping
Ortholog grouping levels (OrthoDB)
Kingdom
Phylum
Class
Order
Family
Ortholog Group Size
Ortholog Group Specificity
Example (LagD):
ABC transporter
ABC transporter ATP binding protein
multidrug ABC transporter permease
bacteriocin cleavage /export ABC transporter
peptidase C39
B
step I
Choose a gene of interest
seed
Find orthologs in other genomes
step II
Assign ortholog groups to neighbours
Kingdom
Phylum
step III
C
Hierarchical data selection
Taxonomic groups of genomes (NCBI)
Bacteria
Firmicutes
Proteobacteria
α-Proteobact.
γ-Proteobact.
Enterobacteriales
Number of genomes
Taxonomic depth
step IV
Select subset of genomes
Proteobacteria
Enterobacteriales
step V
Weight individual genome contribution
-2
-1
seed
+1
+2
x 0.7
x 0.3
x 0.5
x 2
x 1.2
x 0.7
x 1.0
step VI
Calculate weighted frequencies
Weighted Frequency
step VII
Remove statistically insignificant ortholog groups
step VIII
Calculate statistics of conservation
Directional Conservation
100%
75%
50%
25%
0%
25%

2 - Figure 2.1 - Neighbourhood collection workflow

An outline of the workflow used in the data collection for GenCoDB. First, we choose a gene to analyse (the seed) and find its orthologs in other genomes based on a hierarchical clustering sensitivity from OrthoDB. Using the same ortholog assignment strategy, this is repeated for all genes in the seed's neighbourhood. Then we divide the genomes into different subsets based on taxonomic divisions. We calculate the genetic diversity each genome provides to its subset to determine the weight of information a neighbourhood brings to the dataset. Then based on expected conservation frequencies we remove insignificant neighbours resulting in a significantly conserved genomic context. The coloured arrows represent genes on a genome and the different colours represent distinct ortholog groups. The ortholog group of the seed gene is always indicated in red. The shaded boxes behind the steps represent which ortholog group level was assigned to the neighbouring genes (green and blue) or which taxonomic subset of genomes were used to perform the subsequent analysis (orange and yellow)

Gene-centric neighbourhood statistics

To provide comprehensive gene neighbourhood analysis on a broad statistical basis, we focussed our analysis on the 5,487 fully sequenced prokaryotic genomes deployed in OrthoDB v10 (Kriventseva et al. 2019). Given that many traditional analyses of conserved gene neighbourhoods rely on whole genome alignments for which the computation time scales exponentially with the number of genomes (Wolf, Rogozin, Kondrashov, et al. 2001), analysis of such large datasets is not feasible with conventional methods. Thus, to overcome this limitation, GenCoDB considers only local genetic neighbourhoods with a gene-centric approach, allowing for quicker data collection, analysis and visualization, as outlined in Figure 1B. GenCoDB's analysis pipeline starts with the selection of a "seed" gene of interest (Figure 2.1B; step I), and all proteins that are orthologous to it (at a chosen ortholog grouping level as defined by OrthoDB). We exploited the database links between OrthoDB, UniProt and NCBI to identify their respective genomic positions (Figure 2.1B; step II). Then we retrieved the 25 genes up- and downstream for all genes in this selection, recorded their transcription orientation relative to the seed gene and assigned each neighbouring gene with an ortholog group (Figure 2.1; step III). Whenever the link between the three databases was incomplete for a gene (~5% of genes), we assigned the gene with an "unknown" ortholog group thereby maintaining the correct relative positions of up- and downstream genes. For simplicity, in assigning ortholog groups to neighbouring genes, we restricted the possibilities to 5 levels, namely kingdom, phylum, class, order and family (Figure 2.1A). If a gene is not associated with an ortholog group at that level, the closest ortholog group from the next more general taxonomic level is assigned instead. For example, if a gene is missing the ortholog group at the phylum level, the ortholog group from the super phylum is assigned instead. To enable differentiated genetic context analyses at different taxonomic levels, we repeated the downstream analysis for various selections of genomes from different (sub-)taxa, e.g. 'Bacteria' at the kingdom level, 'Proteobacteria' at the phylum level, or 'Enterobacteriales' at the order level (Figure 2.1C). These taxonomic nodes are based on the taxon-definitions of the NCBI database (NCBI Resource Coordinators 2018) and were restricted to those containing a minimum of 50 genomes. In total, GenCoDB analyses genomic context conservation at 89 taxonomic nodes, which not only allows fine-grained studies of gene context conservation in particular taxa, but also enables tracing of gene synteny over evolutionary history (see below). In the following steps we derived gene neighbourhood conservation statistics at different taxonomic levels (Figure 2.1; steps IV-VIII).

2.2 Data correction and normalization

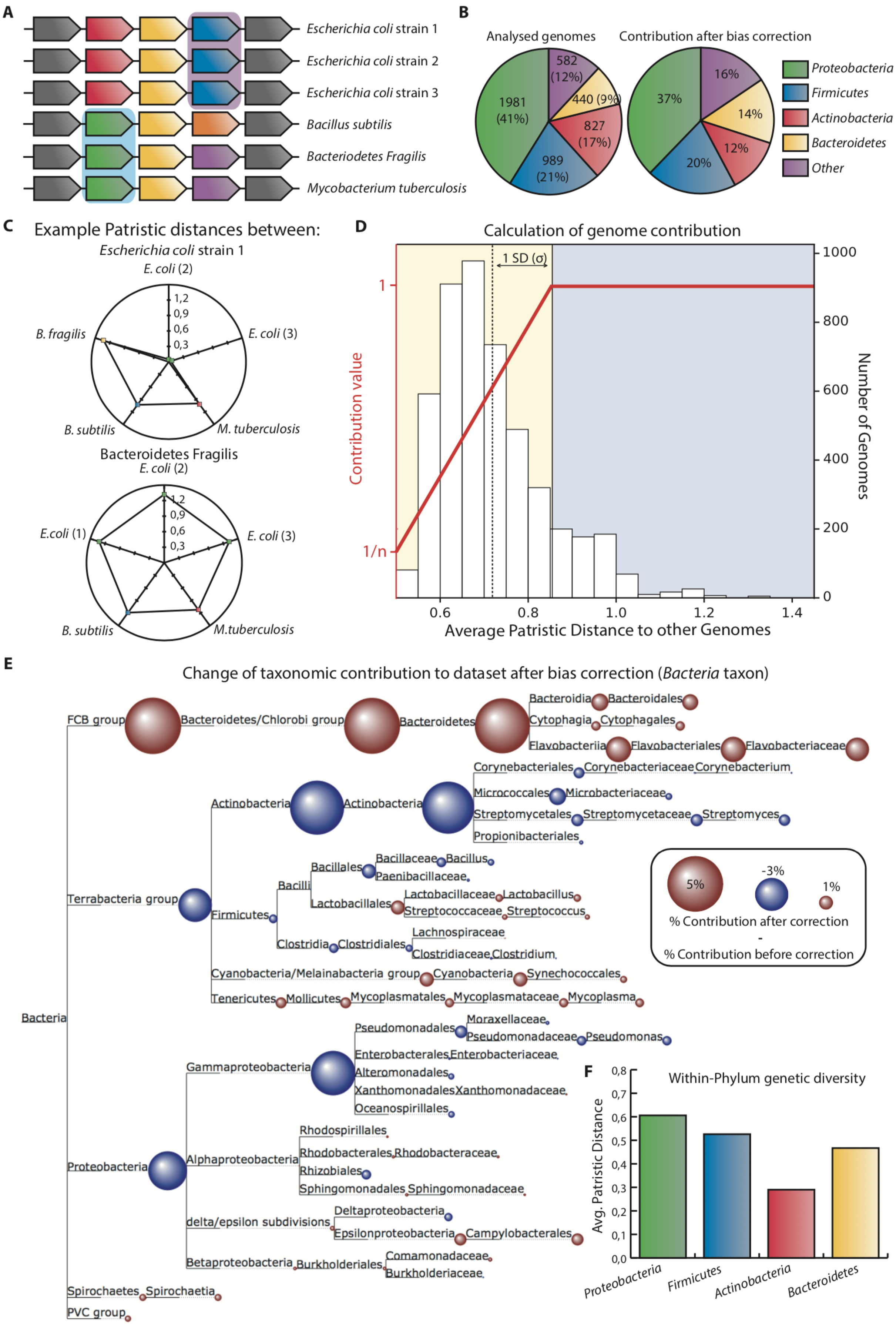
Correction of sampling bias

The published genome sequences of Bacteria are not evenly distributed across the kingdom, with a bias towards species that are: pathogenic to humans, of economic interest, or that are easy to culture in the laboratory. This leads to individual taxa contributing larger fractions of information to a final dataset than others. Therefore, simply counting the absolute number of genomes in which two or more genes are conserved in a particular position does not accurately reflect the true level of conservation, and conversely produces false positive results (Figure 2.2A). For example, if two ortholog groups are both present in two halves of the analysed genomes (see the blue and green genes Figure 2.2A), the high abundance of blue orthologs in closely related *E. coli* sub-strains is far less significant than the high abundance of green orthologs, which are present in far more diverse bacterial species.

Within the underlying dataset of GenCoDB we found indeed an unequal representation of genomes from the four biggest bacterial phyla, with a high abundance of Proteobacteria genomes and a lower abundance of Bacteroidetes (Figure 2.2B). In order to derive meaningful synteny statistics at each taxonomic node, we aimed at correcting for any potential sequencing bias present in the genomes deployed in OrthoDB (Figure 2.1; step IV). To this end we calculated a contribution value for each genome, relative to how distant a species is from closely related species in the dataset. This contribution value overvalues genetically divergent genomes, and devalues genomes which are similar to other species in the dataset. As a measure of evolutionary distance of a given genome, we calculated its average patristic distance (Stuessy and König 2008) - based on a 16S rRNA gene tree (see Supplemental Text) – to its closest neighbours (the 50% of closest genomes) (Figure 2.2C; histogram). We then heuristically assigned a genome contribution value for this genome by linearly interpolating between a minimal and a maximal contribution value with increasing patristic distance (Figure 2.2C; red line). Here, capping the contribution value of individual genomes at a maximal value of 1 prevents a few highly divergent genomes from dominating the conservation statistics, while the minimal contribution value is chosen such that the *n* extremely closely related genomes (those with distances values lower than 10% of the mean - when all distances values are subtracted by minimum distance; see red shaded area Figure 2.2D) are weighted with a value of 1/*n* – effectively treating them as the equivalent of a single genome. As the maximum and minimum thresholds are dependent on the included genomes, we reiterated the calculation of the contribution factor for each genome using the different taxon subsets.

After adjustment we noticed several increases and decreases in the contribution from individual taxa (Figure 2.2E). For example, when applying the adjustment to all genomes (the Bacteria taxon), we observed that the genomes associated to Bacteroidetes and other smaller phyla had a higher contribution to the dataset, as would be expected from the fact that these are fewer genomes that are more distantly related to the other phyla (Figure 2.2B). Interestingly, whilst Proteobacteria and Firmicutes are the major contributors to the dataset, the correction changes their proportional contributions only by 4% and 1%, respectively. This is related to the fact that in this dataset both phyla featured the highest within-phyla genetic diversity of the four biggest phyla (Figure 2.2F), explaining why even after the correction, a large fraction of the dataset is composed of proteobacterial genomes (Figure 2.2F). This also rationalizes the large decrease in contribution from Actinobacteria after the correction (Figure 2.2B), as it has a significantly lower level of genetic diversity in its sequenced genomes (Figure 2.2F). However, while for instance the overall contribution from Proteobacteria changes only slightly, it is noteworthy that the contributions from its individual sub-taxa are not. For example, the contribution of Gammaproteobacteria (containing the two highly sequenced genera, Escherichia and Pseudomonas) decreased after the correction, while the contribution from other sub-taxa, such as the Deltaproteobacteria increased (Figure 2.2E). This bias correction step was recalculated separately at every taxonomic level as the contribution proportion of each sub-taxa varies for each sub-division of the dataset.

Application of this process results in GenCoDB being built on a dataset that contains 5487 bacterial genomes distributed across 89 sub-taxonomic nodes within the bacterial kingdom. The analysis can be centred on over 4 million different ortholog groups, whose neighbourhoods collectively include almost 9 million individual genes. Each ortholog group can be viewed at five different taxonomic levels of ortholog grouping, leading to a total number of almost 1.8 billion different genetic neighbourhoods. To the authors' best knowledge this is the largest and most comprehensive collection of bacterial neighbourhood information.


A
Escherichia coli strain 1
Escherichia coli strain 2
Escherichia coli strain 3
Bacillus subtilis
Bacteriodetes Fragilis
Mycobacterium tuberculosis
B
Analysed genomes
582 (12%)
440 (9%)
827 (17%)
989 (21%)
1981 (41%)
Contribution after bias correction
16%
14%
12%
20%
37%
Proteobacteria
Firmicutes
Actinobacteria
Bacteroidetes
Other
C
Example Patristic distances between:
Escherichia coli strain 1
E. coli (2)
E. coli (3)
M. tuberculosis
B. subtilis
B. fragilis
1,2
0,9
0,6
0,3
Bacteroidetes Fragilis
E. coli (2)
E. coli (3)
M.tuberculosis
B. subtilis
E.coli (1)
D
Calculation of genome contribution
1 SD (σ)
Contribution value
1
1/n
Number of Genomes
0
200
400
600
800
1000
0.6
0.8
1.0
1.2
1.4
Average Patristic Distance to other Genomes
E
Change of taxonomic contribution to dataset after bias correction (Bacteria taxon)
Bacteria
FCB group
Bacteroidetes/Chlorobi group
Bacteroidetes
Bacteroidia
Bacteroidales
Cytophagia
Cytophagales
Flavobacteriia
Flavobacteriales
Flavobacteriaceae
Terrabacteria group
Actinobacteria
Actinobacteria
Corynebacteriales
Corynebacteriaceae
Corynebacterium
Micrococcales
Microbacteriaceae
Streptomycetales
Streptomycetaceae
Streptomyces
Propionibacteriales
Firmicutes
Bacilli
Bacillales
Bacillaceae
Bacillus
Paenibacillaceae
Lactobacillales
Lactobacillaceae
Lactobacillus
Streptococcaceae
Streptococcus
Clostridia
Clostridiales
Lachnospiraceae
Clostridiaceae
Clostridium
Cyanobacteria/Melainabacteria group
Cyanobacteria
Synechococcales
Tenericutes
Mollicutes
Mycoplasmatales
Mycoplasmataceae
Mycoplasma
Proteobacteria
Gammaproteobacteria
Pseudomonadales
Moraxellaceae
Pseudomonadaceae
Pseudomonas
Enterobacterales
Enterobacteriaceae
Alteromonadales
Xanthomonadales
Xanthomonadaceae
Oceanospirillales
Alphaproteobacteria
Rhodospirillales
Rhodobacterales
Rhodobacteraceae
Rhizobiales
Sphingomonadales
Sphingomonadaceae
delta/epsilon subdivisions
Deltaproteobacteria
Epsilonproteobacteria
Campylobacterales
Betaproteobacteria
Burkholderiales
Comamonadaceae
Burkholderiaceae
Spirochaetes
Spirochaetia
PVC group
5%
-3%
1%
% Contribution after correction
-
% Contribution before correction
F
Within-Phylum genetic diversity
Avg. Patristic Distance
0,0
0,1
0,2
0,3
0,4
0,5
0,6
0,7
0,8
Proteobacteria
Firmicutes
Actinobacteria
Bacteroidetes

3 - Figure 2.2 - Scaling of genomic contributions at the Bacteria-taxon

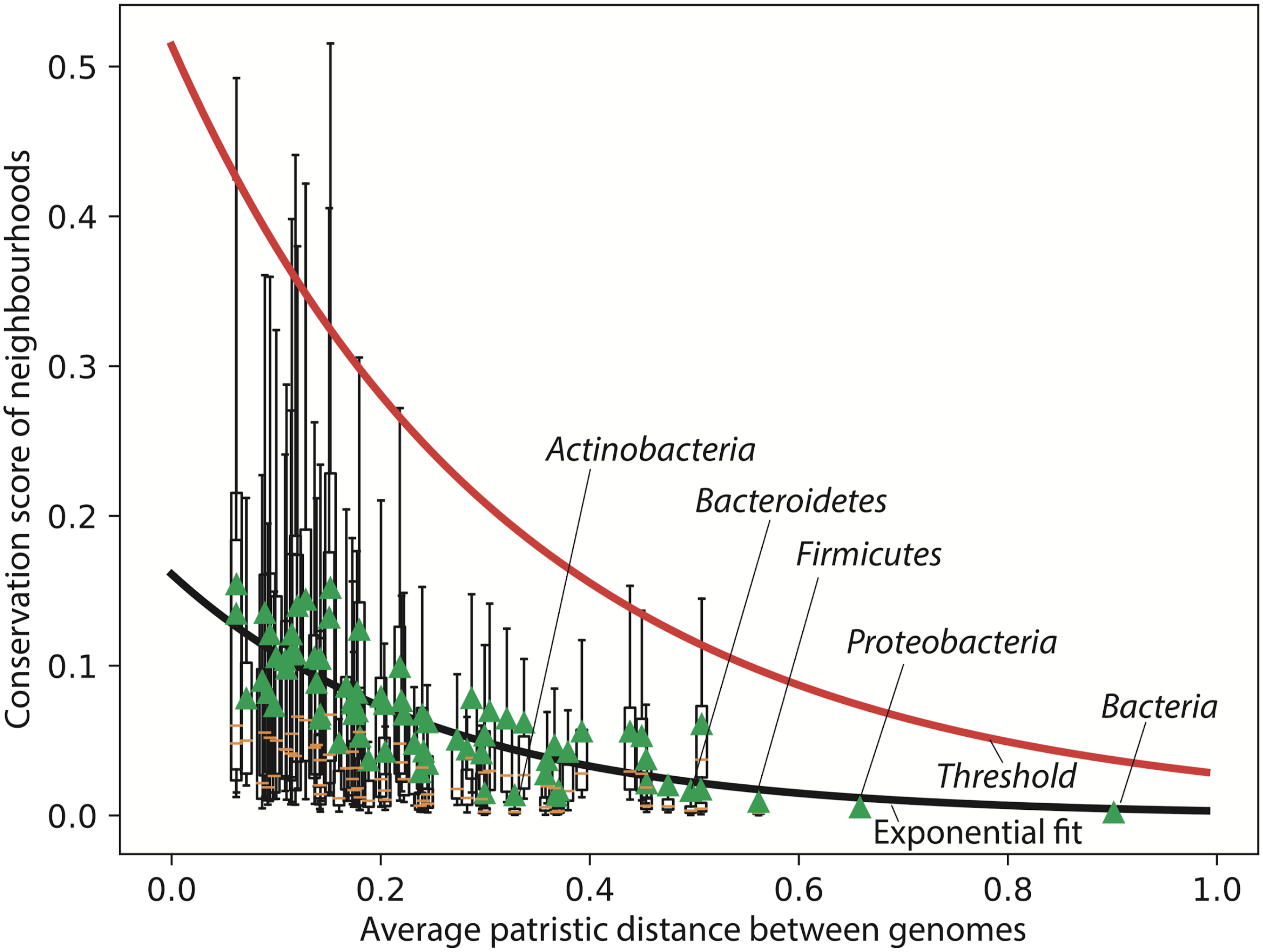
(A) An example genomic neighbourhood highlighting the dangers of not accounting for genomic bias in analysis. Here two orthologs are shown to be conserved in 50% of the genomes, the blue and the green. Whilst in this dataset they are equally conserved as the blue orthologs are all found in E. coli sub-strains it is unlikely this conservation can provide any meaningful biological insight. **(B)** The number of genomes from each phylum in the GenCoDB dataset (left) and the contribution that they make up in the whole dataset (left in brackets and right) before (left) and after bias correction (right). **(C)** An example showing the patristic distances between *E. coli* (top) and *B. fragilis* (bottom) and five other species. The bottom 50% of these values would be taken to calculate the average patristic distance for the species. **(D)** A histogram showing the distribution of average patristic distances of every genome in the dataset. The dotted line indicates the mean of the distribution. Values falling in the scaled region were scaled linearly between the minimum contribution value 1/n and 1 (as represented by the red line). n in this example was 8 genomes. **(E)** A taxonomic tree displaying the change in contribution of each taxa to the dataset at the Bacteria-taxon. Red nodes represent taxa whose contribution is greater after the normalization. Blue nodes represent taxa whose contribution is less. The size of the circles represents the amount of difference. Taxonomic groups with less than 50 genomes are not displayed. **(F)** The average of pairwise patristic distances of all species belonging to the four main phyla.

Determination of significant conservation

Many conserved gene neighbourhoods have been identified in Bacteria, which span both the entire kingdom or are restricted to small bacterial taxa. The number of rearrangement events expected to have occurred since the existence of the shared common ancestor needs to be factored in when considering if the observation of a conserved neighbourhood is significant. If the conservation is higher than what would be expected, this indicates an evolutionary advantage in its maintenance, or contrarily if it is lower than it is most likely present due to chance alone. Determination of this expected rate has been a challenge in this form of analysis. Previous work in this area was limited by lower numbers of available bacterial genomes, and partially solved this issue through the calculation of the probability that two genes occur next to each other on randomly generated, reshuffled genomes (Wolf, Rogozin, Kondrashov, et al. 2001; R. Overbeek et al. 1999). However, the assumption of randomness does not hold true in actual bacterial genome datasets where many genomes are closely related and the chance that one sees two genes next to each other is significantly higher as even distantly related genomes are not fully randomized.

In order to overcome this challenge, GenCoDB determines a threshold for significant context conservation by taking into account the relationship between the genetic diversity within a taxon and the average conservation of neighbourhoods surrounding ortholog groups belonging to them. To this end, for a given set of genomes we first calculated a conservation score of the gene neighbourhood around each ortholog group, by measuring the cumulative frequencies of the top 10 most conserved ortholog groups in the neighbourhood of 25 genes up- and downstream of the seed gene. After normalization a maximum conservation score of 1 is reached if 10 of the neighbouring positions of the seed gene contain a fully conserved ortholog group in 100% of the genomes. The minimum value is 1/(10 x *m*), where *m* is the number of genomes the seed ortholog group is present in, representing the case where each genome has a completely different neighbourhood with no ortholog multiples.

We repeated this process for every taxon, using the average patristic distance of the entire taxon as a measure of the diversity of the group. We then fit an exponential trend line to the relationship between the mean conservation score of each taxon and its genetic diversity (Figure 2.3). To determine a threshold for significant conservation, we calculated an exponential trend line to the standard deviations of the conservation scores for each taxon and added that to the mean fit. Additionally, as neighbourhoods which are strongly conserved and neighbourhoods with a few closely related genomes are included in this calculation, this fit is an overestimation of the true expected conservation value of a gene given a genetic distance of the group. If the conservation of a gene is higher than the value of the trend line at the given genetic distance, it is considered significantly conserved. Using this threshold users can know that what is displayed are relationships that are present higher than one would expect, but if a user does want to change this threshold, either making it more or less sensitive, that is possible.

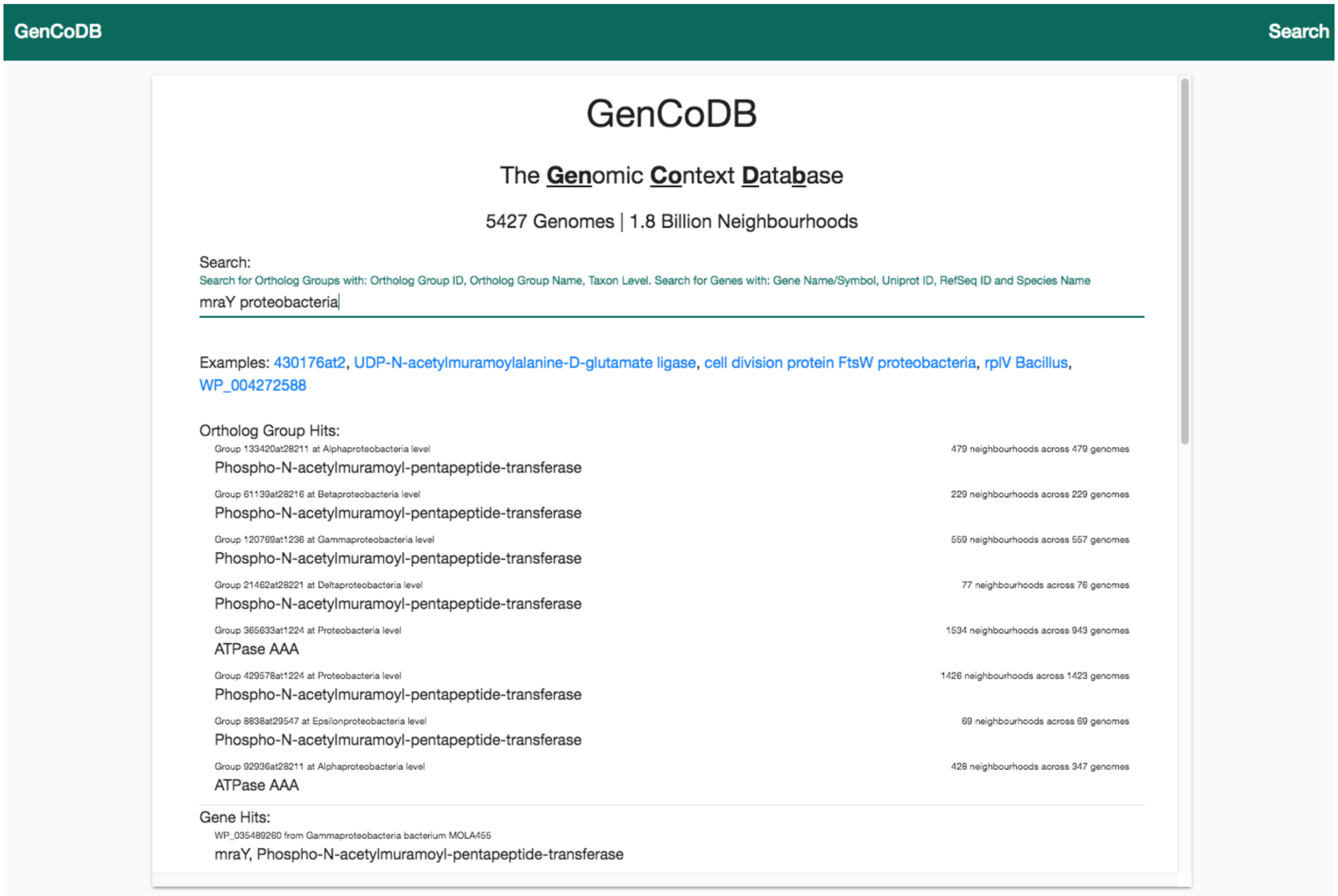


4 - Figure 2.3. Calculation of conservation significance

In this box plot each box represents the average conservation of the 50 most conserved neighbours (conservation score) of each ortholog group at a taxonomic level. The green triangle and yellow line represent the mean and median respectively. The whiskers extend to highest datum within 1.5 interquartile range of the upper quartile and vice versa. Flyers (values greater or less than the whiskers) are not displayed. The black line is an exponential fit of the means and the red line represents a fit to one standard deviation above the mean. Neighbourhoods which have a conservation score higher than the threshold at a taxonomic distance are considered significant.

2.3 GenCoDB user interface

GenCoDB provides multiple views to explore gene conservation information for each bacterial ortholog group from OrthoDB. These views comprise the “Neighbourhood view”, the “Tree view”, the “Genome view”, and one view containing detailed information about the chosen ortholog group. Users can search for ortholog groups by either querying the ortholog group directly via its name or its OrthoDB ID, or by querying the genes belonging to the ortholog group (Figure 2.4). Specific genes can be searched using the gene name/symbol, gene description, UniProt ID or RefSeq ID, which provide a link to a gene-specific page linking to all ortholog groups (at different taxonomic levels) the gene belongs to. To narrow down the search results, the search inputs may also be combined with the taxon or species of interest.



GenCoDB

The **Gen**omic **Co**ntext **D**ata**b**ase

5427 Genomes | 1.8 Billion Neighbourhoods

Search:

Search for Ortholog Groups with: Ortholog Group ID, Ortholog Group Name, Taxon Level. Search for Genes with: Gene Name/Symbol, Uniprot ID, RefSeq ID and Species Name

mraY proteobacteria

Examples: 430176at2, UDP-N-acetylmuramoylalanine-D-glutamate ligase, cell division protein FtsW proteobacteria, rplV Bacillus, WP_004272588

Ortholog Group Hits:

Group 133420at28211 at Alphaproteobacteria level — 479 neighbourhoods across 479 genomes
Phospho-N-acetylmuramoyl-pentapeptide-transferase

Group 61139at28216 at Betaproteobacteria level — 229 neighbourhoods across 229 genomes
Phospho-N-acetylmuramoyl-pentapeptide-transferase

Group 120769at1236 at Gammaproteobacteria level — 559 neighbourhoods across 557 genomes
Phospho-N-acetylmuramoyl-pentapeptide-transferase

Group 21462at28221 at Deltaproteobacteria level — 77 neighbourhoods across 76 genomes
Phospho-N-acetylmuramoyl-pentapeptide-transferase

Group 365633at1224 at Proteobacteria level — 1534 neighbourhoods across 943 genomes
ATPase AAA

Group 429578at1224 at Proteobacteria level — 1426 neighbourhoods across 1423 genomes
Phospho-N-acetylmuramoyl-pentapeptide-transferase

Group 8838at29547 at Epsilonproteobacteria level — 69 neighbourhoods across 69 genomes
Phospho-N-acetylmuramoyl-pentapeptide-transferase

Group 92936at28211 at Alphaproteobacteria level — 428 neighbourhoods across 347 genomes
ATPase AAA

Gene Hits:

WP_035489260 from Gammaproteobacteria bacterium MOLA455
mraY, Phospho-N-acetylmuramoyl-pentapeptide-transferase

5 - Figure 2.4. An example search using GenCoDB.

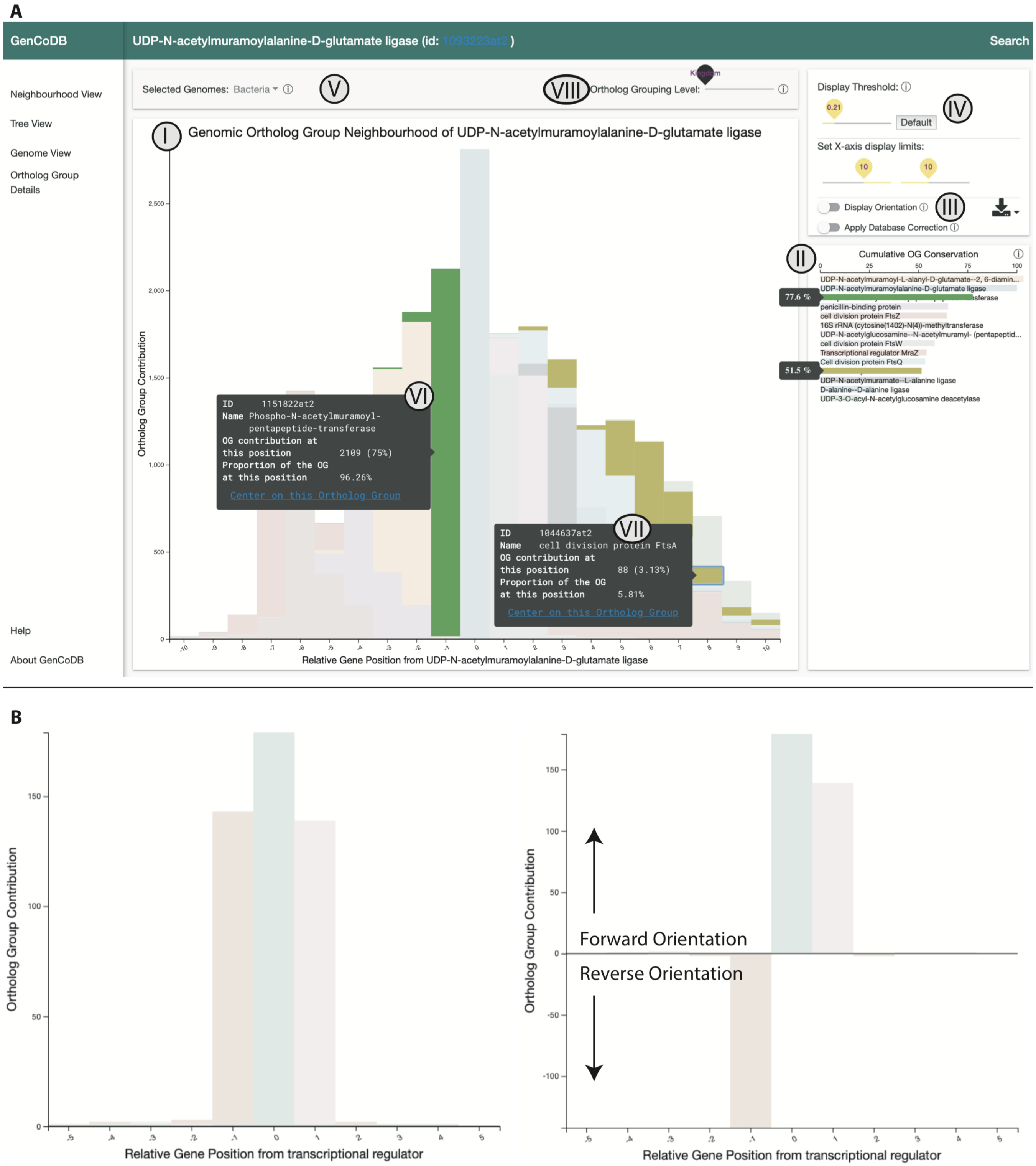
(A) An example displays of GenCoDB being used to search for the *mraY* (UDP-N-acetylmuramoylalanine-D-glitamate ligase) ortholog group at the Proteobacteria level. Both hits to *mraY* containing ortholog groups and specific gene hits are displayed with additional information to inform the user on which group/gene they should use.

Neighbourhood view. One of GenCoDB’s innovations in gene context analysis is the neighbourhood view, which strikes a balance between information content and displaying large amounts of genomic data. The visualized neighbourhood is shown as a stacked histogram representing the abundance and syntactic distance (gene distance) of different ortholog groups relative to a chosen seed gene/ortholog group (Figure 2.5A, I). This captures the advantages of a genome scale viewers (e.g. MicrobesOnline) and summarized neighbourhood viewers (e.g. StringDB), providing detailed information on the variations of neighbours without limiting the number of genomes that can be visualized at once. The neighbourhood view is dynamic in that hovering over a bar will highlight the location of the same ortholog group at other positions and additionally provide a popup containing the statistics of conservation for this group at that position (Figure 2.5A, VI). In an example case, when analysing the genomic neighbourhood of murD (encoding UDP-N-acetylmuramoylalanine-D-glutamate ligase), we observe that there is high conservation of murG (encoding phospho-N-acetylmuramoyl-pentapeptide- transferase) upstream of murD in 75% of the genomes (Figure 2.5A, VI, dark green bar). However, when synteny is not strongly conserved (for example due to insertions/deletions or rearrangements), the bar is separated across multiple columns, such as for the cell division protein ftsZ (Figure 2.5A, VII, light green bars) which at the hovered over area (+8) is present in 3.13% of the genomes there comprising 5.81% of the total ftsZ in this neighbourhood. Therefore, to assist in data comprehension, GenCoDB displays another graph showing the cumulative conservation of each ortholog group in a neighbourhood 25 genes up- and downstream of the seed gene (Figure 2.5A, II), showing that murD and ftsZ are present in 77.6% and 51.5% of neighbourhoods with murD respectively. In addition, gene orientation is very relevant in gene context analysis, for example in predicting the targets of transcriptional regulators, such as for two-component systems, which often regulate divergently transcribed target genes. Therefore, the user may optionally display conservation of genes encoded in the opposite orientation from the seed gene on the negative y-axis. This is exemplified in Figure 2.5B, showing the neighbourhood surrounding a transcriptional regulator (green) of a sigma factor (pink) and a two-component regulator (brown), which is shown to be encoded in the reverse orientation in relation to the other two genes, after the setting has been turned on (Figure 2.5A, III).

As outlined in Figure 2.3, a significance threshold was calculated to determine if the conservation of an ortholog group was above what would be expected by chance. This threshold is set by default and only ortholog groups with a cumulative conservation greater than the threshold are displayed. However, users can manually change the display threshold (Figure 2.5A, IV) to visualize any ortholog group appearing in at least 5% of the selected neighbourhoods. The top 50 most conserved ortholog groups in each selected neighbourhood are assigned a random colour, any additional groups are shaded in grey.

Many ortholog groups have gene members dispersed across the bacterial kingdom, however the composition or presence of a conserved neighbourhood will differ across the bacterial taxonomic tree. To allow users to explore the evolutionary variation within neighbourhoods, GenCoDB provides the option to filter by taxon which genomes are visualised in the histogram. By default, the graph includes the genomic context around the seed gene from all genomes containing the ortholog group. Users can then select from which taxa the genomes should be retrieved from (Figure 2.5A, V), thereby allowing for taxon-specific genomic conservation analyses. Importantly, while navigating through different taxonomic levels, GenCoDB automatically adjusts the default significance threshold according to Figure 2.3, in order to accommodate for a different level of genetic diversity in the different genome subsets chosen by the user.

In addition, the user may choose the ortholog grouping level at which the neighbourhood is visualized (Figure 2.5A, VIII). This changes the ortholog groups that are assigned to neighbour genes, but does not change the genomes which are included in the analysis. There is not always a correct or clear choice for which level of ortholog grouping a user should use, as it depends on their research question, gene of interest and their needs. For example, at more specific ortholog grouping levels it is possible that a neighbour is split between two ortholog groups and neither is above the significance threshold and therefore not displayed, even though at a higher grouping level they would be combined and subsequently shown as significantly conserved (Figure 2.1A). Conversely if a less specific grouping is used, the annotations of the group are more general and provide less insightful information, thereby obscuring the conservation of interesting processes in a gene neighbourhood).



A
GenCoDB
UDP-N-acetylmuramoylalanine-D-glutamate ligase (id: 1093223at2)
Search
Neighbourhood View
Tree View
Genome View
Ortholog Group Details
Help
About GenCoDB
Selected Genomes: Bacteria ⓘ
V
VIII
Ortholog Grouping Level:
Kingdom
ⓘ
I
Genomic Ortholog Group Neighbourhood of UDP-N-acetylmuramoylalanine-D-glutamate ligase
Ortholog Group Contribution
2,500
2,000
1,500
1,000
500
0
-10
-9
-8
-7
-6
-5
-4
-3
-2
-1
0
1
2
3
4
5
6
7
8
9
10
Relative Gene Position from UDP-N-acetylmuramoylalanine-D-glutamate ligase
VI
VII
Display Threshold: ⓘ
0.21
Default
IV
Set X-axis display limits:
10
10
Display Orientation ⓘ
III
Apply Database Correction ⓘ
II
Cumulative OG Conservation ⓘ
0
25
50
75
100
77.6 %
51.5 %
UDP-N-acetylmuramoyl-L-alanyl-D-glutamate--2, 6-diamin...
UDP-N-acetylmuramoylalanine-D-glutamate ligase
...sferase
penicillin-binding protein
cell division protein FtsZ
16S rRNA (cytosine(1402)-N(4))-methyltransferase
UDP-N-acetylglucosamine--N-acetylmuramyl- (pentapeptid...
cell division protein FtsW
Transcriptional regulator MraZ
Cell division protein FtsQ
UDP-N-acetylmuramate--L-alanine ligase
D-alanine--D-alanine ligase
UDP-3-O-acyl-N-acetylglucosamine deacetylase
B
Ortholog Group Contribution
150
100
50
0
-5
-4
-3
-2
-1
0
1
2
3
4
5
Relative Gene Position from transcriptional regulator
Ortholog Group Contribution
150
100
50
0
-50
-100
Forward Orientation
Reverse Orientation
-5
-4
-3
-2
-1
0
1
2
3
4
5
Relative Gene Position from transcriptional regulator

```
ID      1151822at2
Name Phospho-N-acetylmuramoyl-
     pentapeptide-transferase
OG contribution at
this position           2109 (75%)
Proportion of the OG
at this position        96.26%

  Center on this Ortholog Group
```

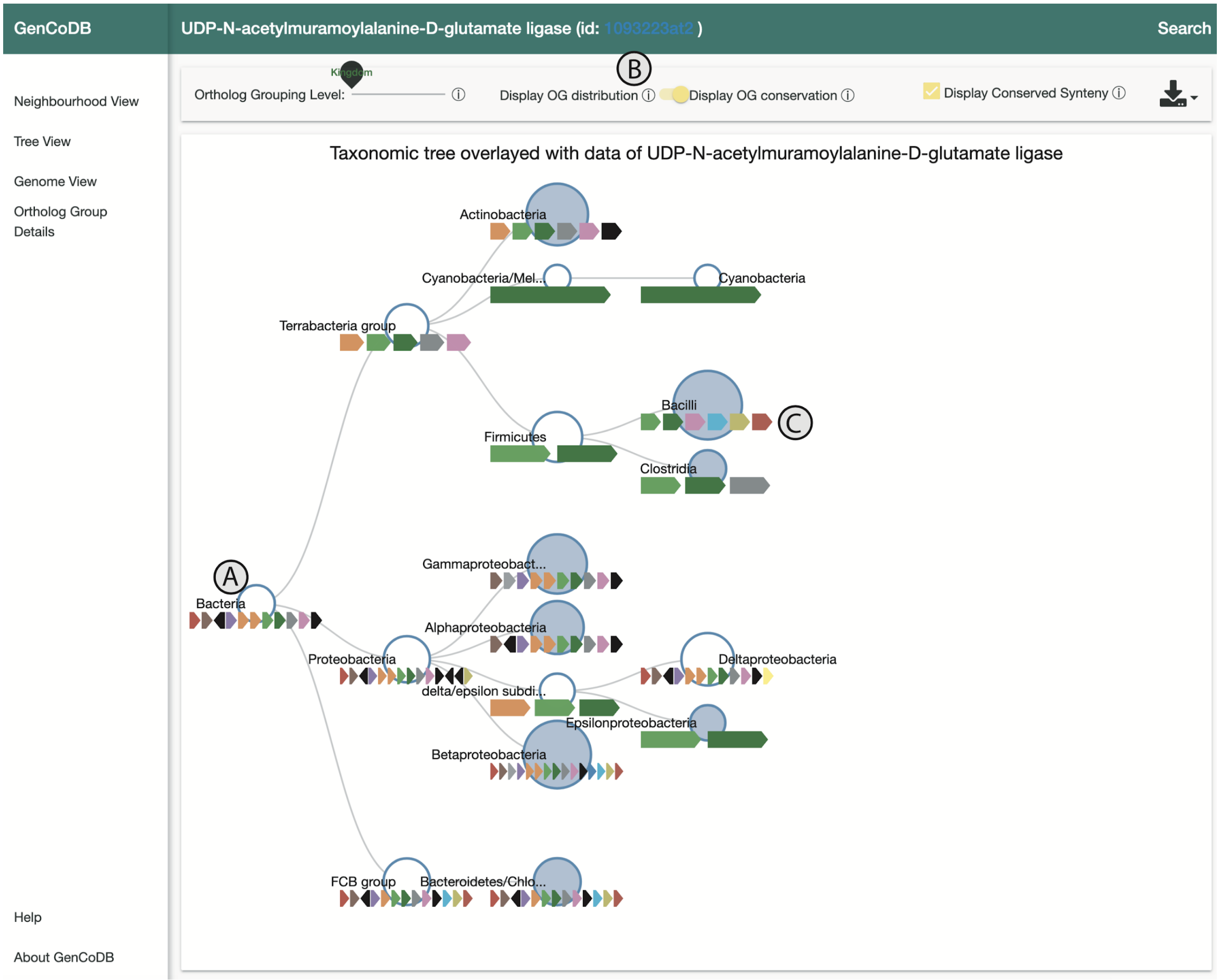
```
ID      1044637at2
Name    cell division protein FtsA
OG contribution at
this position           88 (3.13%)
Proportion of the OG
at this position        5.81%

  Center on this Ortholog Group
```

6 - Figure 2.5. An example display of the Neighbourhood View

(A) An example display of *murD* (UDP-N-acetylmuramoylalanine-D-glutamate ligase) in the neighbourhood view. (1) The histogram represents the frequency of each ortholog group appearing at a gene distance from *murD*. Each bar can be hovered over to provide statistics on the conservation of that group at that position relative to the seed gene (*murD* in this case), see (VI and VII). (II) The bar plot on the right displays the total conservation of each ortholog group in the neighbourhood. (III) The ability to display genes which are not co-oriented with the seed gene in the negative y-axis can be toggled on (See B). The bias correction can also be toggled, swapping between raw genome frequencies and the calculated contribution values. (IV) Here the threshold for which ortholog groups are displayed based on their total conservation in the neighbourhood. By default, it is set at the level we calculated is significant for the taxonomic group the user is using. (V) Here the user can quickly change the subset of genomes they are looking at. This does not change the ortholog group sensitivity. (VIII) Here the user can change the ortholog group assignments of the neighbouring genes, this does not change the underlying genome subset, and the user would need to select their seed ortholog group at the correct taxonomic division for that. **(B)** An example display of the neighbourhood of a transcriptional regulator in the default state (left), and when orientation is being shown (right).

Tree view. As alluded to in the previous section, the context surrounding genes can vary drastically in different taxa and can provide insightful information about the evolution or function of gene clusters. For example, the conservation of a particular gene cluster in only a subset of bacterial taxa can indicate an important physiological function of this gene arrangement, and tracing the addition/removal of genes to/from the cluster over evolutionary time may reveal interesting correlations with physiological behaviour of bacteria. To analyse such events, GenCoDB provides the tree view, which summarizes the genetic context of the neighbourhood view and projects it on a phylogenetic tree for the underlying taxonomic groups of genomes (Figure 2.6A), see the supplementary text for details. In this view, the user can inspect the most conserved synteny at each taxonomic node - very similar to the functionality provided by StringDB (Snel et al. 2000). Here, the conserved synteny is defined by the most conserved genes at a particular position, if the conservation is above the significance threshold for this taxonomic group. In addition, the user can toggle to display either the conservation of the seed gene in each taxonomic group (i.e., the fraction of species containing the seed gene in that group), or the conservation score of the neighbourhood in the taxonomic group (Figure 2.6A). Here, the conservation score is calculated as the average cumulative conservation of the top 10 most conserved genes in the neighbourhood, which can simply be interpreted as the area of the bars found in the neighbourhood view. Using this statistic it is possible to quickly identify taxonomic groups with a highly conserved neighbourhood around the chosen seed gene, allowing the user to further focus on these taxonomic in the other views. This also helps users to identify possible functions of a conserved cluster, as it facilitates determining which taxa maintain the observed clustering and which have lost it.

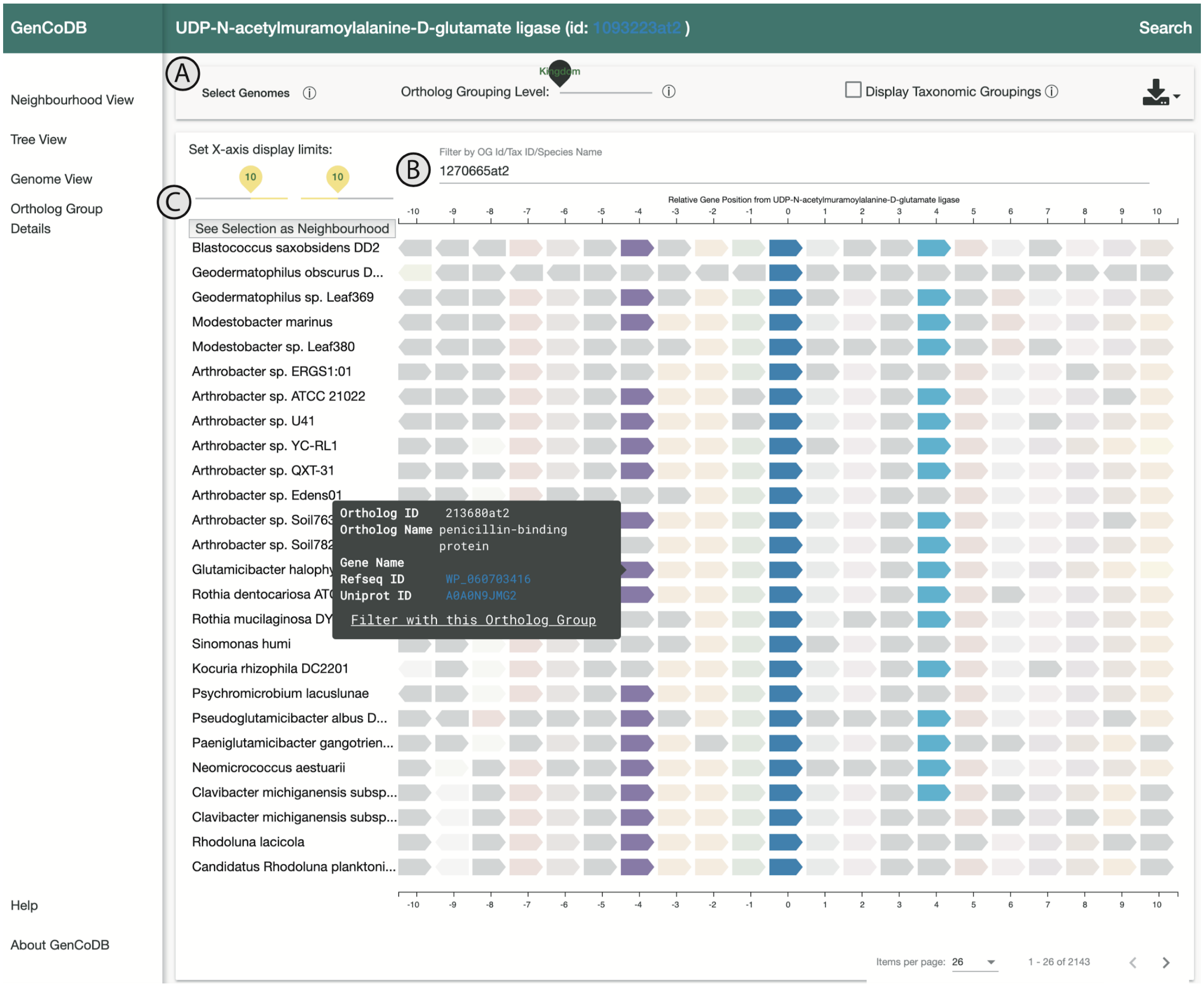
For both display and convenience purposes only the highest three taxonomic levels are displayed by default, however users are able to show or hide a node's descendants by clicking on them. Taxa with less than 50 representatives of the ortholog group are not displayed in this mode. If a more detailed analysis is required for the genomic context at a particular taxonomic node, users can navigate directly from the tree view to the neighbourhood view via a link shown in the tooltip associated with each taxonomic node.



7 - Figure 2.6 - Example display of the Tree View

An example display of *murD* (UDP-N-acetylmuramoylalanine-D-glutamate ligase; dark green arrow) in the tree view. (A) The circles represent the conservation score of the ortholog group at that taxonomic level. The dark circles mean the circle can be clicked to reveal children taxonomic groups. (B) The user can decide if the size of the circle should show the conservation score or the number of neighbourhoods in the taxonomic group with this ortholog group (C) The arrows below the circles represent the most conserved synteny surrounding the seed gene at the taxonomic level and can be hovered over to see what ortholog groups they are. A gene will only be considered part of the synteny if its conservation is above the significance threshold at that position. The seed gene is always shown in dark green.

Genome view. This view provides two main functions for GenCoDB: Firstly, users are able to inspect the raw gene context around their chosen ortholog group in individual genomes. Therefore, if a novel gene neighbourhood is found in other views, in this view it is possible to determine which genomes contain the gene cluster, and which species have different genomic rearrangements. The second function is that it allows for a customized selection of genomes (beyond taxonomic membership), which can then be subjected to further statistical analysis (Figure 2.7A). For instance, users may choose their own subsets of bacteria, such as human pathogens or flagellated bacteria, for display of the genomic context. Additionally, users are able to easily filter the view to only show genomes containing a combination of particular ortholog groups in their neighbourhood (Figure 2.7B). This combinatorial search quickly narrows down the displayed genomes to only those containing the cluster of interest, allowing to user to discover other co-occurrences which may not have been as apparent in the whole dataset. Importantly, once all the desired genomes have been selected either by filtering by the presence of ortholog groups or the selection of genomes or a combination of the two, these can be exported to be visualized in the neighbourhood view (Figure 2.7C). However, as the process of calculating the contribution bias from species and the significance threshold is computationally too expensive, these statistical corrections are not available in the neighbourhood view of custom genome selections.


GenCoDB
UDP-N-acetylmuramoylalanine-D-glutamate ligase (id: 1093223at2)
Search
Neighbourhood View
Tree View
Genome View
Ortholog Group Details
Help
About GenCoDB
A
B
C
Select Genomes ⓘ
Ortholog Grouping Level:
Kingdom
ⓘ
□ Display Taxonomic Groupings ⓘ
Set X-axis display limits:
10
10
See Selection as Neighbourhood
Relative Gene Position from UDP-N-acetylmuramoylalanine-D-glutamate ligase
-10
-9
-8
-7
-6
-5
-4
-3
-2
-1
0
1
2
3
4
5
6
7
8
9
10
Blastococcus saxobsidens DD2
Geodermatophilus obscurus D...
Geodermatophilus sp. Leaf369
Modestobacter marinus
Modestobacter sp. Leaf380
Arthrobacter sp. ERGS1:01
Arthrobacter sp. ATCC 21022
Arthrobacter sp. U41
Arthrobacter sp. YC-RL1
Arthrobacter sp. QXT-31
Arthrobacter sp. Edens01
Arthrobacter sp. Soil763
Arthrobacter sp. Soil782
Glutamicibacter halophy
Rothia dentocariosa ATC
Rothia mucilaginosa DY
Sinomonas humi
Kocuria rhizophila DC2201
Psychromicrobium lacuslunae
Pseudoglutamicibacter albus D...
Paeniglutamicibacter gangotrien...
Neomicrococcus aestuarii
Clavibacter michiganensis subsp...
Clavibacter michiganensis subsp...
Rhodoluna lacicola
Candidatus Rhodoluna planktoni...
Filter with this Ortholog Group
Items per page: 26
1 - 26 of 2143

Filter by OG Id/Tax ID/Species Name: 1270665at2

```
Ortholog ID   213680at2
Ortholog Name penicillin-binding
              protein
Gene Name
Refseq ID     WP_060703416
Uniprot ID    A0A0N9JMG2
```

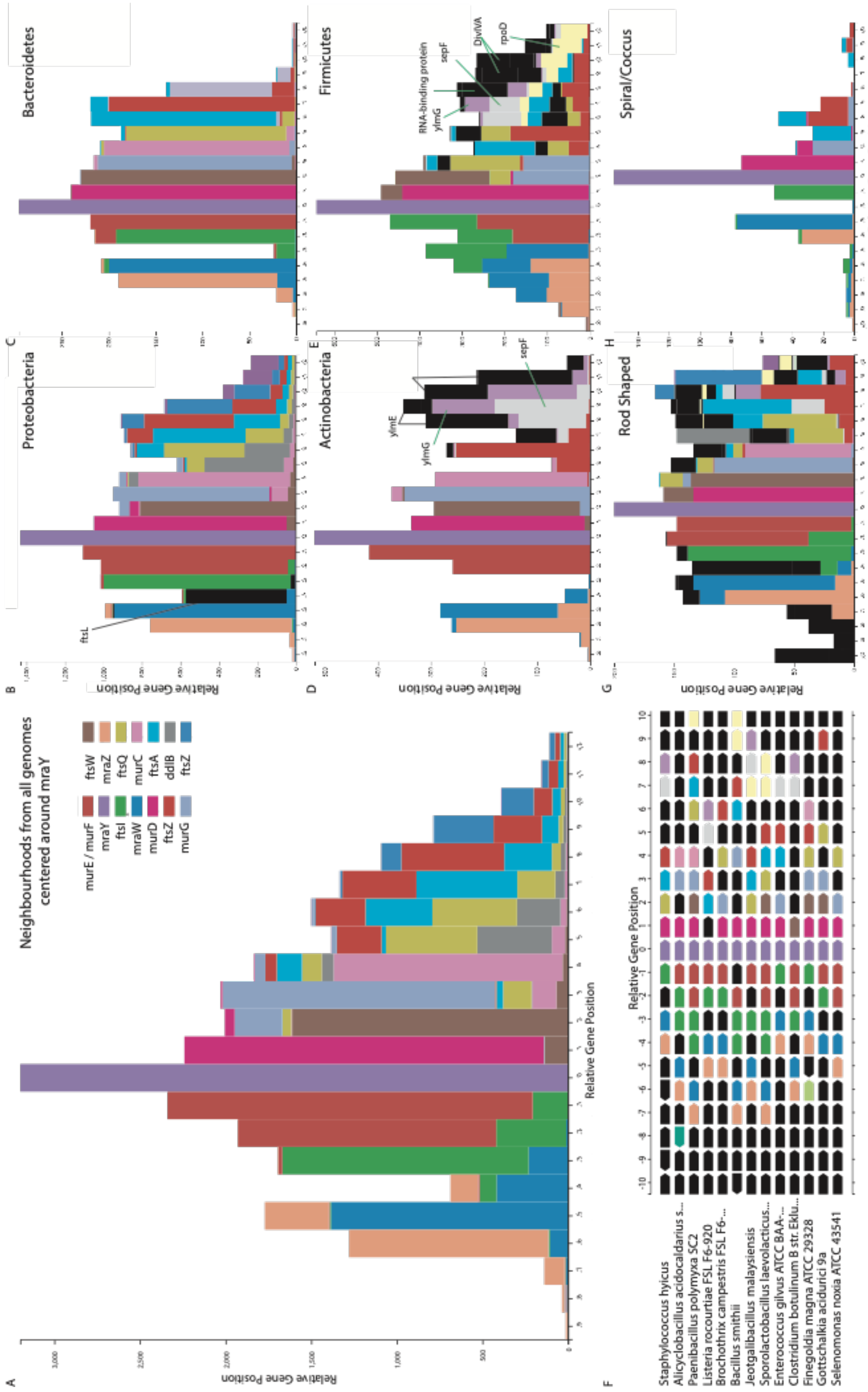
8 - Figure 2.7 - An example display of the Genome View

In this example, the ortholog group of a *murD* is displayed in the genome view. (A) Here the users can select which taxa or species they would like to see displayed in this view (or for later export to the neighbourhood view) (B) Here the user can filter to only include neighbourhoods of the seed gene that contain the specified ortholog group, in this case the light blue gene (they may be found ±25 up or downstream). One gene is hovered over showing various identifiers for that gene and links to the represpective databases. (C) Once users are happy with their filtered selection of neighbours, they can export it to the neighbourhood view to see it as a quantitative histogram.

Data availability. Through the user interface every graph is available to download, both in *.svg and *.png formats, allowing the effortless generation of publication-quality graphics. Both the neighbourhood and genome view allow for download of the raw data in comma separated value (csv) format. In particular, the *.csv files available from the neighbourhood view contains a row for each ortholog group in the displayed neighbourhood, with columns containing the frequency of that ortholog group appearing in the 25 up- and downstream positions surrounding the seed gene. The genome view produces a *.csv file which has a row for each selected genome and in the columns the ortholog group assigned to the genes in the 25 up- and downstream positions surrounding the seed gene. Both of these *.csv files reflect the settings selected in the user interface, including the database correction, genome selection and orientation options (a + or – will be placed before ortholog group IDs to signify relative orientation to the seed gene). These *.csv files allow for the reproduction of the graphs with other visualization strategies or for further downstream analyses.

2.4 Application of GenCoDB to the division cell wall cluster

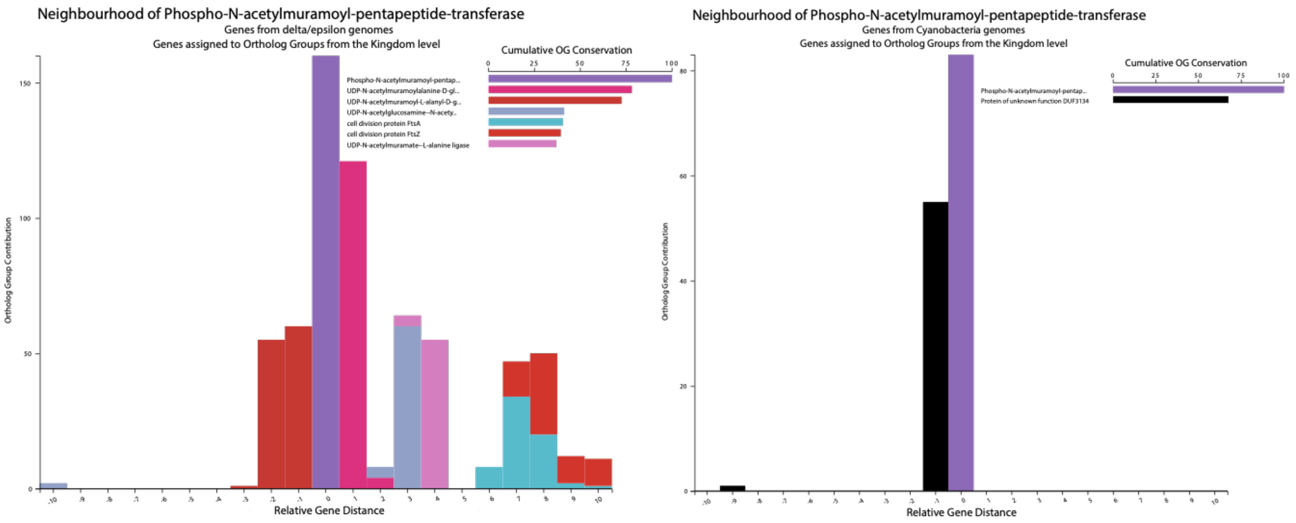
To test the functionality of GenCoDB we applied it to the task of analysing the DCW cluster, a gene cluster which is conserved throughout the bacterial kingdom, containing genes responsible for various steps in the synthesis of cell wall precursors and cell division. The cluster has been well described for various model organisms, such as *E. coli* (M. Vicente, Gomez, and Ayala 1998), *B. subtilis (Real and Henriques 2006)* and *N. gonorrhoeae (Francis et al. 2000)*, typically containing around 15 genes that mostly belong to the *mur* family (responsible for cell wall precursor synthesis) and *fts* family (responsible for cell division)(Tamames et al. 2001). We first focused on *mraY*, one of the genes that appears in the centre of the cluster in many model organisms. Indeed, when considering all genomes, we found 13 other genes that were significantly conserved in this neighbourhood (meaning a cumulative co-conservation with *mraY* over 22% with the lowest at 27.7% being D-alanine--D-alanine ligase) with minor tapering of conservation for genes which are positioned further away from *mraY*, a difference between (Figure 2.8A). We also observed that co-direction orientation is highly conserved within the cluster that there were no cases where gene inversion of these orthologs had occurred. This points to operon organised transcription and indeed it has already been observed in several species that many of these genes are shown to be regulated by the same promoters (M. Vicente, Gomez, and Ayala 1998; Francis et al. 2000; de la Fuente, Palacios, and Vicente 2001). Strikingly conservation of the first 10 genes seem to have the position of the gene relatively strongly conserved however the last 4 genes whilst overall similarly conserved as the other genes, have very variable positions suggesting several insertion and deletion events within the right side of the cluster.


A
Neighbourhoods from all genomes centered around mraY
murE / murF
mraY
ftsI
mraW
murD
ftsZ
murG
ftsW
mraZ
ftsQ
murC
ftsA
ddlB
ftsZ
Relative Gene Position
B
Proteobacteria
ftsL
C
Bacteroidetes
D
Actinobacteria
ylmE
ylmG
sepF
E
Firmicutes
RNA-binding protein
ylmG
sepF
DivIVA
rpoD
F
Staphylococcus hyicus
Alicyclobacillus acidocaldarius s...
Paenibacillus polymyxa SC2
Listeria rocourtiae FSL F6-920
Brochothrix campestris FSL F6-...
Bacillus smithii
Jeotgalibacillus malaysiensis
Sporolactobacillus laevolacticus...
Enterococcus gilvus ATCC BAA-...
Clostridium botulinum B str. Eklu...
Finegoldia magna ATCC 29328
Gottschalkia acidurici 9a
Selenomonas noxia ATCC 43541
Relative Gene Position
-10 -9 -8 -7 -6 -5 -4 -3 -2 -1 0 1 2 3 4 5 6 7 8 9 10
G
Rod Shaped
H
Spiral/Coccus

9 - Figure 2.8 – Analysis of the DCW cluster in GenC

(A) The neighbourhood of *mraY* including all genomes in which *mraY* is found. **(B-E)** The neighbourhood of mraY with genomes from only Proteobacteria, Bacteroidetes, Actinobacteria and Firmicutes respectively. The legend for the coloured bars is found in A. Black bars represent ortholog groups that were not in the top 50 most conserved groups when considering all genomes and certain groups have been labelled for convenience. **(F)** A selection of genomes from firmicutes showing the distribution of genes around mraY. Each arrow represents a gene, and the colour the assigned ortholog group. Purple represents mraY and the other colours match the legend in A and B with slight opacity. Black arrows represent genes that are not displayed in the histogram view as they are not considered significantly conserved. **(G,H)** The neighbourhood of mraY with a custom selection of genomes either of rod-shaped bacteria **(G)** or cocci and spiral shaped bacteria **(H)**. The colour of the bars is consistent with the legend from A and B.

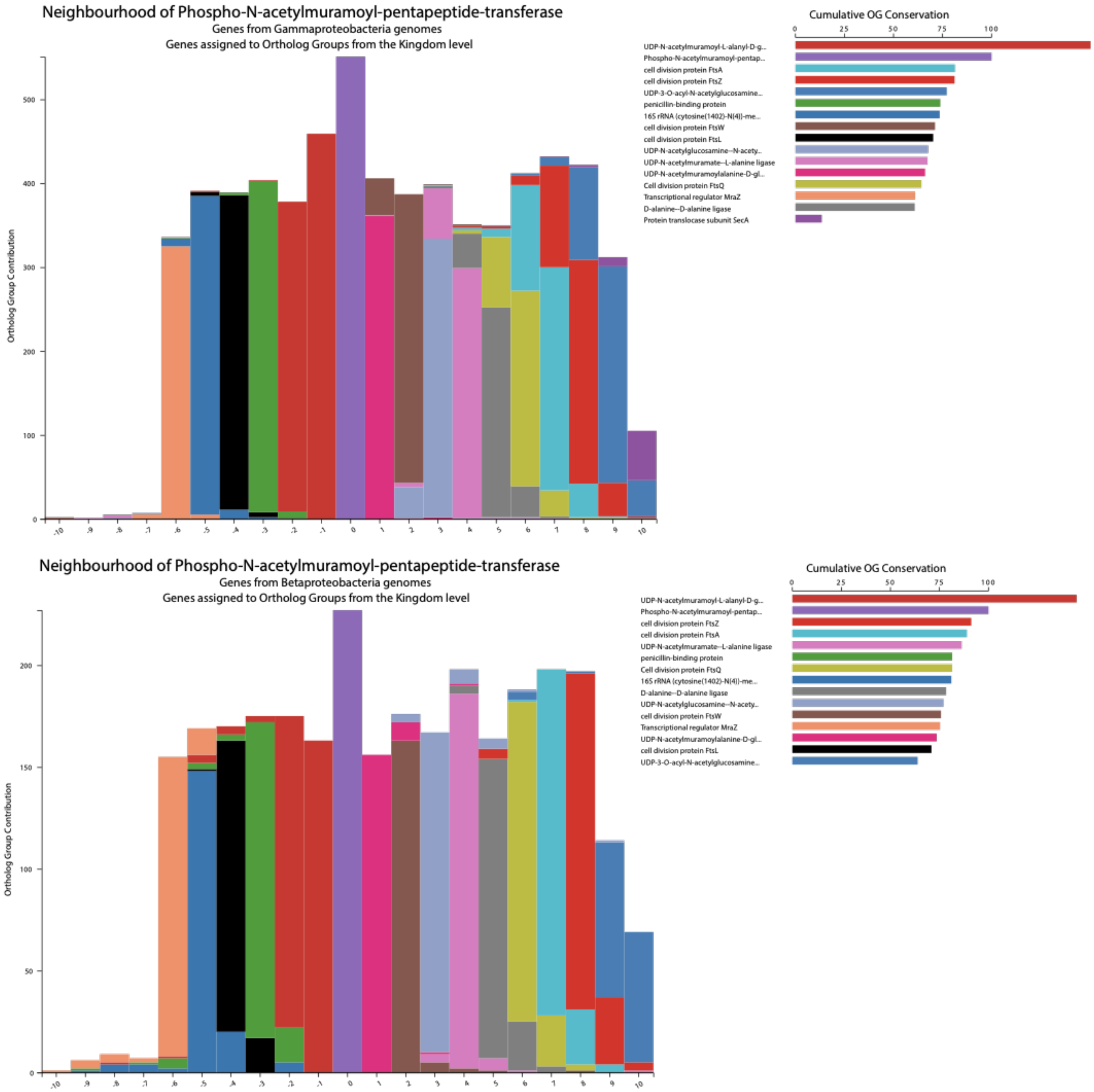
Curious about these rearrangements, we wanted to see how these changes were distributed across the bacteria kingdom and if they were localized to particular taxa. Viewing *mraY* in the tree view we see that the conservation score is strikingly lower in the Cyanobacteria and delta/epsilon subdivisions of Proteobacteria (17.37 and 19.56 respectively) (Figure 2.5). A closer inspection in the neighbourhood view confirmed that in many genomes of these sub taxa the neighbourhood around *mraY* was gone (Figure 2.9). We also noticed that the conserved synteny whilst mostly similar across the different taxa was much smaller in firmicutes than the other phylum even through the conservation scores were relatively similar and in fact slightly higher than proteobacteria or bacteroidetes which have large conserved syntenies (30.78 vs 28.41 and 28.27 respectively). Therefore, to investigate why this was the case we restricted the number of genomes to the main 4 phyla in our database: Proteobacteria, Actinobacteria, Firmicutes and Bacteroidetes.



10 - Figure 2.9 – Loss of conserved genomic neighbourhoods of mraY in some taxa

The genomic context surrounding *mraY* from delta/epsilon Proteobacteria genomes (left) and Cyanobacteria genomes (right). Height of the bar represents the conservation of an ortholog group, with colours signifying the different ortholog groups. Legends are found in the top right-hand corner of each histogram.

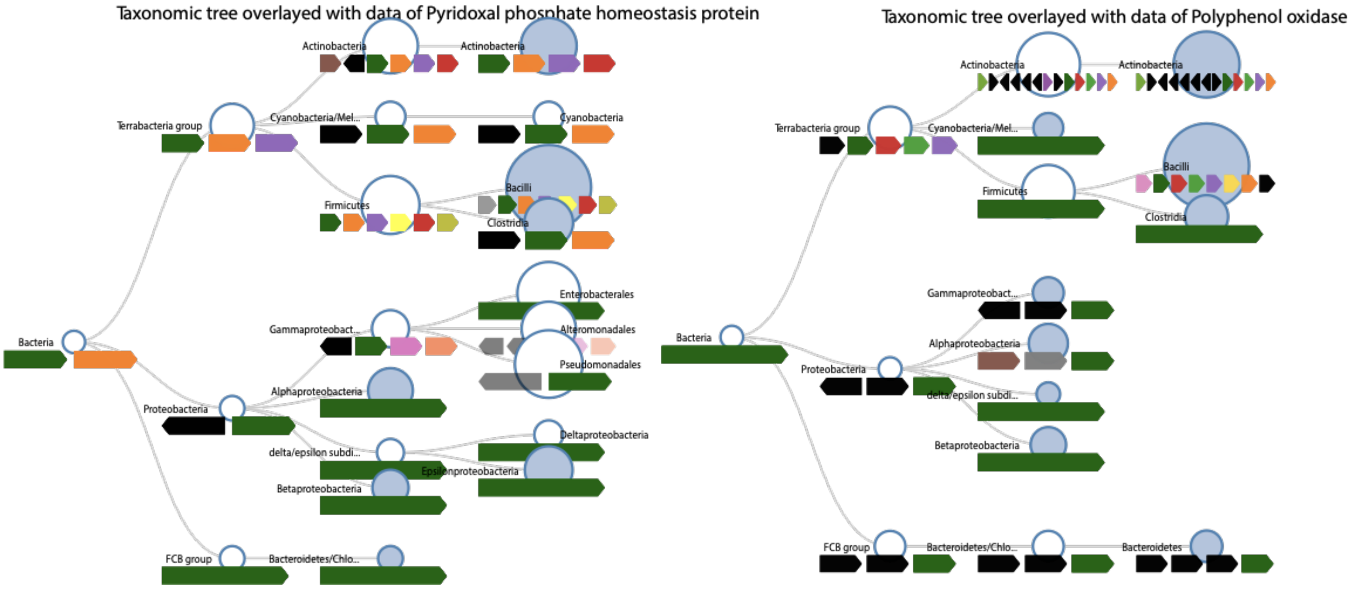
In Proteobacteria and Bacteroidetes we only see slight disruption and when we explored further down the taxonomic tree there was no to little disruption in the gamma and beta proteobacteria (Figure 2.8B and C, Figure 2.10). In addition to the core DCW genes, many Proteobacteria genomes contain ddlB and *ftsA* (Figure 2.8B), being co-conserved with *mraY* in 56% and 67.5% of neighbourhoods, whereas Bacteroidetes contains a glutamyl-tRNA aminotransferase (*yqeY*) at 45.9% (Figure 2.8C). However, in actinobacteria we see the association of 5 additional ortholog groups, namely a pyrodoxal phosphate homeostasis protein (*ylmE*) (56.6%), a polyphenol oxidase (*ylmD*) (47.6% - below the significance threshold), Cell division protein SepF (69.7%), an uncharacterized membrane protein (*ylmG*) (58.4%) and a divIVA domain containing protein (Figure 2.8C) (72.6%). Furthermore, these were accompanied with the loss of *ftsA* (Figure 2.8C).



11 - Figure 2.10 – Strongly conserved genomic neighbourhoods of mraY in some Gamma/Betaproteobacteria

The genomic context surrounding mraY from Gammaproteobacteria genomes (left) and Betaproteobacteria genomes (right). Height of the bar represents the conservation of an ortholog group, with colours signifying the different ortholog groups. Legends are found in the top right-hand corner of each histogram.

Cell wall synthesis and division occurs differently in actinobacteria, as unlike other rod-shaped bacteria they elongate their cell wall from their poles and not laterally along the cell length (reviewed in (Pamela et al. n.d.)). Also, in contrast to the other phylum, a major player in the divisome, FtsZ, is not essential for growth for Actinobacteria (McCormick et al. 1994), therefore it is unsurprising there would be difference in the division cell wall cluster. The biological function of many of the newly introduced genes to the cluster have yet to be determined however given that in our tool we see them associated with the DCW cluster, a role in cell division or septum formation seems likely. Indeed, SepF and *divIVA* which do not have close homologs in non-terrabacteria genomes, have been shown to be crucial in Z-ring formation leading to division (Hamoen et al. 2006). *ylmE* and *ylmD* have orthologs in the majority of bacterial species however according to our tool, they do not have significant conservation partners except in terrabacteria (Figure 2.11). In *Streptomyces venezuelae* these two genes were deleted and there was no observable impact on growth rate, septum formation or sporulation (Santos-Beneit et al. 2017), however this does not preclude a role in cell wall synthesis and in non-laboratory conditions. For instance in *E.coli* (not from Actinobacteria) *yfiH* (the *ylmD* homolog) was shown to be involved in preventing non-canonical amino-acids from being incorporated into the peptide chain in place of L-alanine (Parveen and Reddy 2017). As actinobacteria do not have D-alanine--D-alanine ligase (*ddlB*) in their DCWcluster, perhaps *ylmD* provides a complementary function. We also found evidence that uncharacterized membrane protein (*ymlF*) may have a small role in cell division as a knock out mutant for this gene in *Streptococcus pneumoniae* was shown to have thinner septums and increased numbers of tetrads and diplococci suggesting incomplete division (Fadda et al. 2003), and in chloroplasts, the *ylmG* ortholog when overexpressed, impaired chloroplast division and distribution of the chloroplast nucleoids(Kabeya et al. 2010). Therefore, one could extrapolate, in the absence of such confirmatory literature the functional predictive power conserved neighbourhoods provide and how this could be used to help functionality characterise currently under researched ortholog groups and the role players in novel phenotypes unique to certain clades.



12 - Figure 2.11 – Evolution of conserved neighbourhoods of *ylmE and ylmD*

The evolutionary history of the genomic neighbourhoods of *ylmE* (left) and *ylmD* (right). The circles represent the conservation score of the ortholog group at that taxonomic level. The arrows below the circles represent the most conserved synteny surrounding the seed gene at the taxonomic level. A gene will only be considered part of the synteny if its conservation is above the significance threshold at that position. The seed gene is always shown in dark green.

In Firmicutes there is significant disruption both on 3' and 5' ends of the cluster (Figure 2.8E). Upstream of *mraY* the same genes are conserved as found in the other phyla but it appears the order of the genes is greatly intermingled, however closer inspection of the genomes in the genome view shows that the relative order of these 5 genes remains the same but they have been randomly interspaced with other genetic elements (Figure 2.8.F). Downstream of *mraY* there is a loss of *murC* and *ddlB* and in some cases the addition of the same ortholog groups that appeared in the actinobacteria neighbourhood, such as the DivIVA domain-containing protein and the uncharacterized membrane protein (*ylmG*) (Figure 2.8E). Given Actinobacteria and Firmicutes share a more recent common ancestor compared to the other phyla it is reasonable that they share common rearrangements in this cluster and signifies these changes likely occurred before the division of terrabacteria. Two new additions to this cluster were the RNA polymerase sigma factor RpoD, and an RNA-binding protein (Figure 2.8E). RpoD is the housekeeping sigma factor active during exponential growth and up-regulates genes associated with fast growth such as translation associated proteins (Ozaki et al. 1991). As cell wall synthesis and cell division also occurs at a faster rate during high growth rates and less required during other phases perhaps associating this cluster with this sigma factor may allow for faster response times to changes in nutritional availability.

Here, we would like to mention two important notes that highlight the benefits of the flexible customization of GenCoDB. Firstly, at the default parameters, the *murE* and *murF* genes belong to the same ortholog group. This would occur if the protein sequences of these two genes are very similar to one another and it has been shown that murE and murF despite not having over high sequence similarity have highly conserved motif regions and most likely diverged from a recent common ancestor (Bouhss et al. 1997). Without prior knowledge of the cluster it is not clear these are two separate genes with two functions, however by adjusting the ortholog grouping level to a lower level, in this case phylum, *murE* and *murF* cluster into distinct groups. Secondly *ftsL* is only found in the Proteobacteria despite it being highly conserved in this cluster in all phyla. This is because despite being recognized as the same gene, the differences in the sequences of *ftsL* cluster the proteins separately even in the least sensitive of ortholog group levels.

Using these observations and then connecting them with literature confirms the power of genetic context analysis for hypothesis generating however it can also be used to provide confirmatory evidence of research questions. Tamames(2001) found that the conservation of this cluster was correlated with cell morphology, specifically rod-shaped cells. To test this observation we looked at the neighbourhood of mraY in the known rod shaped Bacilli and other filamentous bacteria (e.g Actinomyces, Clostridium, Enterobacter) and compared that to a neighbourhood of non-rod bacteria (e.g coccoids such as Streptococcus, Enteroccocus and Neisseria bacteria and spiral shaped bacteria from Helicobacter, Campylobacter and Leptospira). Here we see that in the rod bacteria there is a strong conserved neighbourhood around *mraY* however in spiral and coccoidal bacteria there is no to little conservation surrounding *mraY*, confirming what was reported by Tamames(2001) (Figure 2.8G and H). Given this striking difference it is tempting to propose that if through random rearrangement events the DCW cluster is broken, the interplay between the different proteins of this cluster is demolished and the coordination required to form a rod shaped cell wall is lost. Alternatively, this evidence could suggest that the selective pressures that maintain the DCW cluster are only present in rod shaped bacteria, and if they lose this cell morphology through disruptions in other parts of the genome, reshuffling of the DCWcluster would then be permitted. Further in lab investigation especially looking at the organisms which do not follow the trend of being rod shaped with a DCW cluster would be required in order to to tease these two alternatives apart.

2.5 Comparison with current tools

GenCoDB holds a unique position amongst other bacterial genome comparison tools as the way for non-bioinformatic trained scientists to get quantitative gene neighbourhood data. However, how does it compare to the many other tools in its other aspects? An important distinction between these tools is how they determine orthologs between different genomes. Of the four surveyed tools: StringDB, MicrobesOnline, JGI's IMG/M tool (Chen et al. 2019) and the EFI-Genome Neighbourhood Tool(Gerlt et al. 2015), a mix of COG and PFAM identities are used to group the genomes (Table 2.1). GenCoDB is the only that is built on the OrthoDB classifications. With the use of COG and PFAM groupings comes the loss of stratification in the clustering sensitivities. MicrobesOnline and JGI provide some customization with two different forms of grouping in their visualization with MicrobesOnline giving the option of MicrobesOnline Ortholog groups instead of COGs to provide increased sensitivity. GenCoDB gives the option of 5 different levels of clustering sensitivity for neighbours. Alternatively, COG and PFAM ortholog group identifiers support significantly more genomes than OrthoDB classifications and therefore more genomes can be included in the analysis. However, only JGI takes advantage of this and contains more genomes in their published dataset than GenCoDB, and only with the latest updates is StringDB close to matching GenCoDB in the number of species (Table 2.1). The gene neighbourhood sections of MicrobesOnline, JGI and the EFI-Genome Neighbourhood Tool are specialized in showing genomes aligned to a by a chosen seed ortholog group which is emulated in the Genome view in GenCoDB. As previously discussed, this form of visualisation is restricted by the number of genomes that can be displayed at one time, StringDB overcomes this by summarising the neighbourhoods and only displaying the most conserved gene synteny of each taxon, as seen in the Tree view of GenCoDB. Another important factor for genome comparison research is finding where and how often two or more genes co-occur with one another. GenCoDB provides this function through the genome view, and of the other four surveyed tools only StringDB and JGI provide this functionality, however in StringDB it is limited to only genes that are considered highly networked.

Of great importance is how these tools facilitate the analysis of genomic context. GenCoDB is the only tool which performs a correction for sampling bias in the genomes present in their dataset facilitating the interpretation of the data and the formation of conclusions. By measuring, quantifying and reporting the strength and quality of conservation reduces the requirement for users to perform further analysis downstream to get interpretable data. With the quantitative data that GenCoDB provides, users can do more than hypothesis generation but directly ask impactful and relevant research questions.

Comparison of genome neighbourhood comparison webtools

	GenCoDB	**StringDB**	**Microbes Online**	**JGI**	**EFI-Genome Neighbourhood Tool**
Number of Bacterial Genomes	5487	4445	1752	14088	UniProt Database
Ortholog Grouping	OrthoDB (5 different levels)	COG	COG/Microbes Online Ortholog Groups	COG and PFAM	PFAM
View at Genome Level	Yes	No	Yes	Up to 40 at a time	Yes
Synteny evolution	Yes	Yes	No	No	No
Co-occurrence	Yes	Yes*	No	Yes	No
Handles Genome Bias	Yes	No	No	No	No
Quantitative Neighbourhood	Yes	No	No	No	No
Suitability for non-bioinformaticians	+++	+++	+++	++	+

1 - Table 2.1 - Comparison of features from different Bacteria genome comparison tools

A comparison of the features and data of GenCoDB to other popular Bacteria genome comparison tools. Ortholog grouping refers both to how ortholog groups are classified and how many levels of sensitivity are provided. View at Genome level referred to the capability to seeing genome by genome which genes are next to the seed gene. Synteny evolution refers to the function to quickly see how the average gene order has changed between different taxonomic phylum. Co-occurrence asks if the tool allows you to find out exactly which genomes contain two or more genes co-localized. Qualitative neighbourhood refers to if the tool provides statistics on exactly how conserved different ortholog groups are with each other. '+' - represents a qualitative score, with more +'s meaning a higher evaluation. * - Only for high networked genes

2.7 Summary

Studying the genetic context of bacterial genes has proven to be an extremely useful tool for geneticists in classifying and identifying gene function, interaction partners and regulation networks. Of special interest are the cases where the context is conserved across multiple genomes, signifying a selective advantage of this genomic arrangement. The ubiquitous need for these data is evidenced by the multiple online platforms which provide genome browsers. However, current tools are not well equipped for the increasingly large number of bacterial genomes that are being made available and only provide relatively simple analyses, such as the co-occurrence of two ortholog groups and completely precludes more complex analyses including phenotype correlation, phylogenetic analyses and evolution studies. Additionally, as sequencing databases inflate so does the impact of sequencing bias, which results in the increased likelihood of false positive conclusions from what was previously a wellspring of insight.

To address this, we created GenCoDB, a highly interactive and dynamic database aiming to provide the scientific community, a hub for sourcing genomic context conservation data in bacteria. The data are a collation of over 5000 bacterial genomes whose genes are aligned through their ortholog groupings in order to calculate quantitative conservation statistics of their gene neighbours. By utilizing the hierarchical ortholog classifications by orthoDB (Kriventseva et al. 2019), users are not restricted to only one ortholog group per gene, allowing for adjustable sensitivity in finding orthologs, which is especially important when analysing highly abundant but diversified proteins such as DNA bindings proteins. Users can analyse their chosen ortholog group through three distinct "views": The genome view provides a look at the raw alignment of their gene/ortholog group of interest. Of particular utility is that the displayed genomes can be filtered based on both taxa/subtaxa membership and the co-occurrence of other ortholog groups in the neighbourhood. This selection of genomes can be exported to the neighbourhood view, which displays the conservation of ortholog groups at relative distance from a "seed" gene via a stacked bar plot, providing quantitative statistics of conservation and allowing the visualisation of thousands of genomic contexts simultaneously. Here, the statistics can be adjusted to account for the aforementioned sequencing bias and by default only significantly conserved ortholog groups are displayed. In the tree view the context surrounding a gene can be tracked at different taxonomic levels providing an overview of the average synteny seen in each taxon and giving clues, e.g., as to when genomic rearrangements may have occurred and in which groups of organisms the genomic context remains conserved

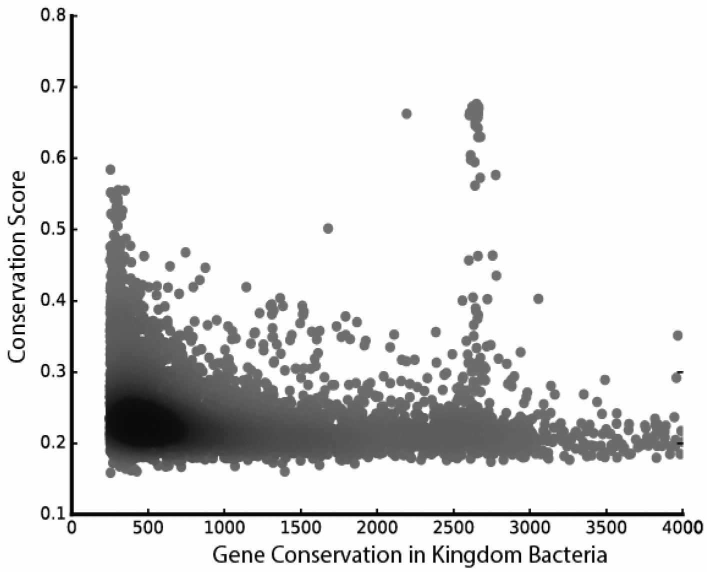
3. Bioinformatic analysis of bacterial gene cluster evolution

In this chapter we take the large amounts of quantitative genomic context data that we produced for GenCoDB and use it to better understand the evolution and dynamics of gene clustering in bacteria. By focusing these data, we identify conserved gene clusters at a high taxonomic resolution allowing us to track their evolution over evolutionary time. Using this state-of-the-art list of gene clusters, we can test the long-standing hypotheses in the field to determine the selective pressures that maintain gene clusters. We also propose a new model for the emergence of operons within clusters and that it does not provide a long-term selective pressure in maintaining gene clusters.

3.1 Analysis of genomic contexts

<u>The relationship between gene conservation and neighbourhood conservation</u>

In Chapter 2 we measured the average conservation found in neighbourhoods from different bacterial taxa (Figure 2.3). There we saw that neighbourhood conservation increased as the genetic diversity (and therefore possible rearrangement events) decreased. Contrary to this, we know that essential genes are often broadly conserved and are less likely to be involved in rearrangement events. Therefore, we set out to observe the impact of these two opposing forces using our neighbourhood dataset by comparing conservation around each gene with their taxonomic dispersal. Indeed, both forces were strongly visible in our dataset. We saw a peak of high neighbourhood conservation for both genes that are conserved in only a small number of genomes and in genes that are broadly conserved (Figure 3.1). As we would expect, these genes were enriched in essential processes or known to be found in conserved gene clusters and included ribosomal genes, ATP synthases and flagella components. Interestingly, we also observed a great number of neighbourhoods between these extremes, those that are more conserved than the average neighbourhood of a gene with similar conservation across the bacterial kingdom (Figure 3.1). Strikingly, most of these outliers are previously non-cluster associated processes thereby highlighting the depth and power our large genome dataset and method have in identifying new conserved neighbourhoods.

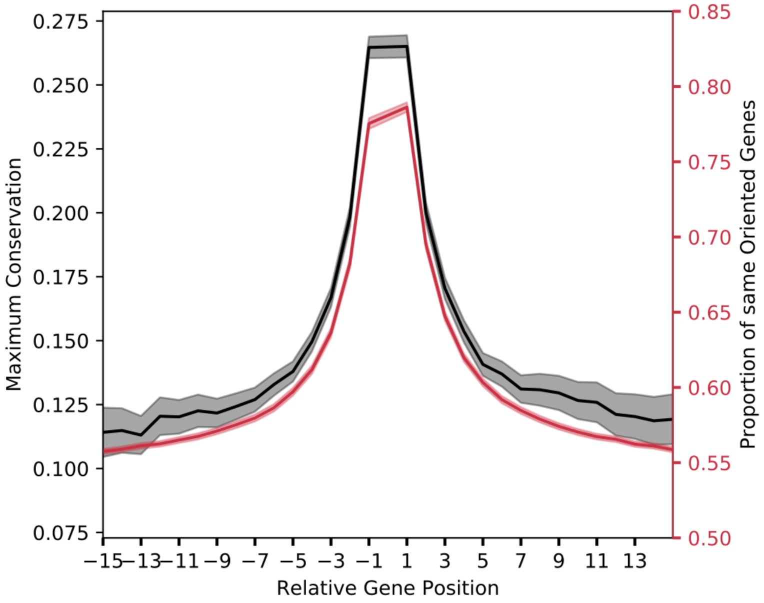


<u>13 - Figure 3.1 - The relationship between conservation of a gene and its genomic neighbourhood</u>

Each point represents a unique ortholog group (gene) at the Bacteria taxonomic level. The conservation score is calculated as the average neighbourhood conservation of the top 50 most conserved neighbour-genes. The gene conservation is a measure of the number of genomes the gene is present in. The contribution of each genome is normalized based on how genetically distinct species was in the dataset. Here broadly conserved genes such as ribosomal proteins found in nearly every organism will have gene conservation values ~2700. Values higher than this represent ortholog groups that have multiple occurrences in genomes and are often highly duplicated genes. The shade of the points represents the density of points at the location.

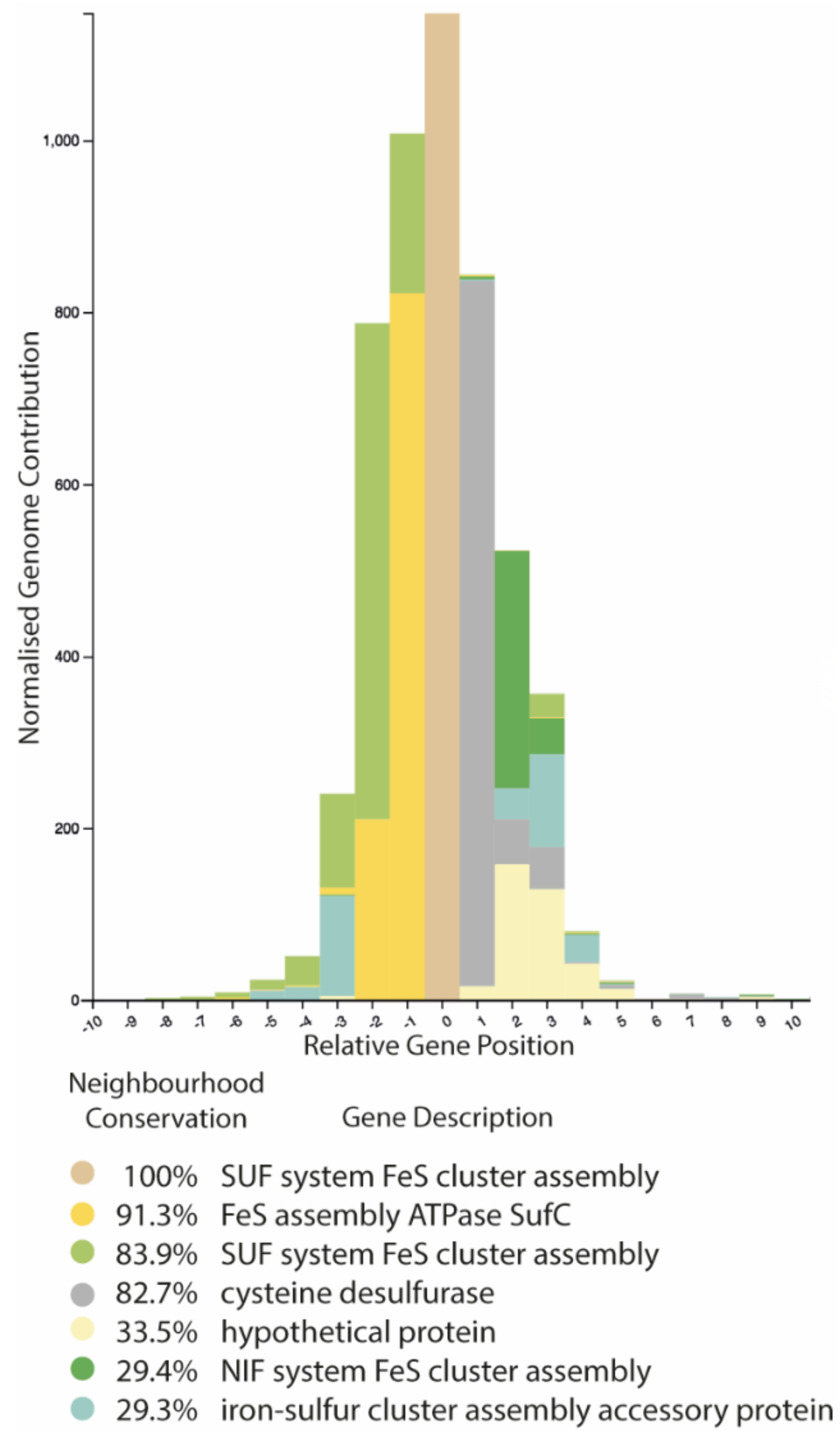
Neighbourhood conservation degrades rapidly with size at a gene scale

The greater the spacing between genes on a genome, the more statistically probable a rearrangement event will occur between them resulting in a separation of their neighbourhoods. Our dataset currently excludes physical chromosome distance but instead uses the number of genes as a measurement of distance. The majority of previous work uses a nucleotide scale to judge rearrangement frequency, therefore to understand the impact of using a gene scale for our dataset we measured neighbourhood decay at a gene scale. To this end, we examined the neighbourhood of each ortholog group containing at least 250 genomes. For each neighbourhood we measured the maximal conservation of the genes found x positions up- and downstream of the original ortholog group relative to the conservation of the ortholog group (positional conservation) (see Figure 3.3 for a visual representation). As expected, we observed a negative correlation between neighbour conservation and gene distance (Figure 3.2). The rate at which this decrease occurred was striking as conservation deteriorated exponentially as a function of distance. We observed no significant bias towards up- or downstream neighbours in relation to conservation. From this information we can state that conserved gene clusters comprised of more than three genes appear to be evolutionarily un-favored (Figure 3.2). This also details an important consideration when looking at the significance of co-localization of two genes next to each other, as we observed an increase of 15% in the conservation of a directly neighbouring gene. Therefore, gene pairs should require an equivalent increase in co-conservation to be considered significant (Figure 3.2). For future analysis we will consider clusters consisting of three or more genes to reduce the impact this distance bias has on our observation.



14 - Figure 3.2 - Conservation of genes and orientation with gene distance

The neighbourhood statistics of ortholog groups which were at least partially conserved (>250 genomes). Position 0 on the x-axis represents the location of the genes belonging to the ortholog group. The black line shows the average of highest positional conservation values for each neighbour at each relative position. The red line represents the average proportion of genes which are in the same orientation as the seed gene at each position. The shared areas (grey, pale red) represent the standard error.


Normalised Genome Contribution
1,000
800
600
400
200
0
-10
-9
-8
-7
-6
-5
-4
-3
-2
-1
0
1
2
3
4
5
6
7
8
9
10
Relative Gene Position
Neighbourhood Conservation
Gene Description
100% SUF system FeS cluster assembly
91.3% FeS assembly ATPase SufC
83.9% SUF system FeS cluster assembly
82.7% cysteine desulfurase
33.5% hypothetical protein
29.4% NIF system FeS cluster assembly
29.3% iron-sulfur cluster assembly accessory protein

15 - Figure 3.3 - Example gene neighbourhood of *sufD*

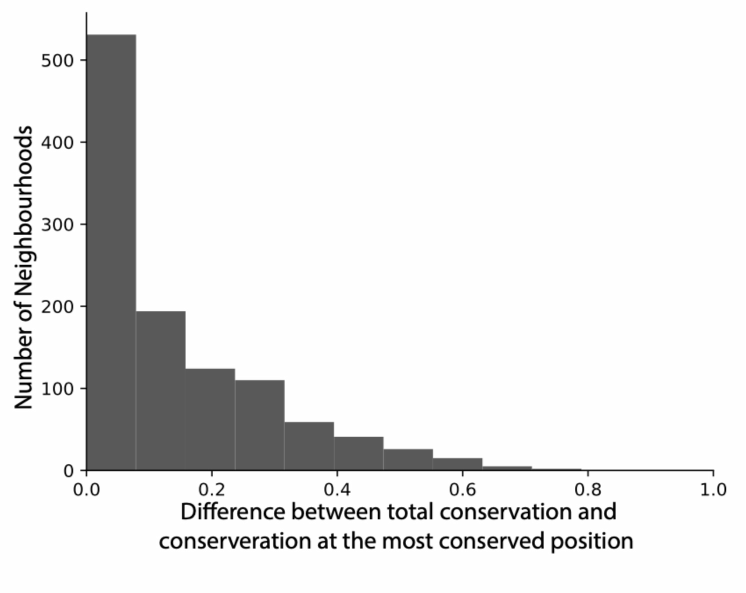
An example representation of a conserved gene neighbourhood centred on *sufD* (Light brown bar in the centre). The colours represent different neighbouring ortholog groups. The legend below the plot matches the colour to the ortholog group. Genome contribution is normalized based on the relatedness to other species in the dataset. The positional conservation is identified by the height of the bars per x-axis position relative to the centred gene (*sufD*) Neighbourhood conservation is the sum of these bars per colour. The neighbourhood conservation value is the sum of positional conservation proportions for each ortholog group is displayed in the legend as a percentage of the countered genes genome contribution. Neighbouring ortholog groups which were less than 5% present (relative to the seed gene) were removed for clarity.

Orientation is tightly regulated in gene neighbourhoods with no external forces

In the introduction we outlined several plausible selective forces that could maintain a gene cluster, one of which being co-regulation of multiple genes through co-transcription as a polycistronic transcript (an operon). This requires that all genes that are part of the cluster are co-oriented. Therefore, we tracked how orientation is biased in our neighbourhoods. If there is no association between two neighbouring genes orientations (therefore random) we would expect that on average co-orientation would occur randomly in 50% of cases. Contrary to this, we observed a bias towards maintaining a similar orientation amongst nearby genes (Figure 3.2). We also saw that the average orientation trended to a value greater than 50% as distance increased. This is likely caused by the orientation bias found in bacterial where genes are preferentially located on the leading strand to avoid collisions with DNA replication machinery (Rocha and Danchin 2003) Unexpectedly, a strong correlation between the conservation of gene neighbourhoods and their orientation was revealed potentially alluding to orientation being a key element in conserved gene neighbourhoods (Figure 3.2). This tight correlation may also suggest that the orientation of gene clusters does not impact genes outside the neighbourhood which regress to the "randomized" genome average.

Short-range synteny is conserved in gene clusters

Due to the constraints of previous alignment methods in identifying conserved gene neighbourhoods, the existence of contexts in which gene content is conserved but gene order is highly malleable are undetectable. Our dataset was generated independent of alignments (excluding the seed gene) and therefore perfectly suited to analyse how often neighbourhoods are shuffled, those which maintain gene content - but in a different synteny. To do this we compared the highest conserved gene in the neighbourhood to its highest positional conservation in each neighbourhood. To ensure we were only capturing neighbourhoods that were conserved in the first place, only ortholog groups which had a neighbour with a neighbourhood conservation of over 20% were considered. Rearrangements were found to happen very infrequently, as neighbourhood conservation and highest positional conservation were often identical, meaning the genes did not change relative position (Figure 3.4). This means insertions, deletions and shuffling must not occur frequently within neighbourhoods. This demonstrates that synteny is critical to the function of gene clusters and that the measured high rate of mutation to gene order is lower in conserved gene neighbourhoods.



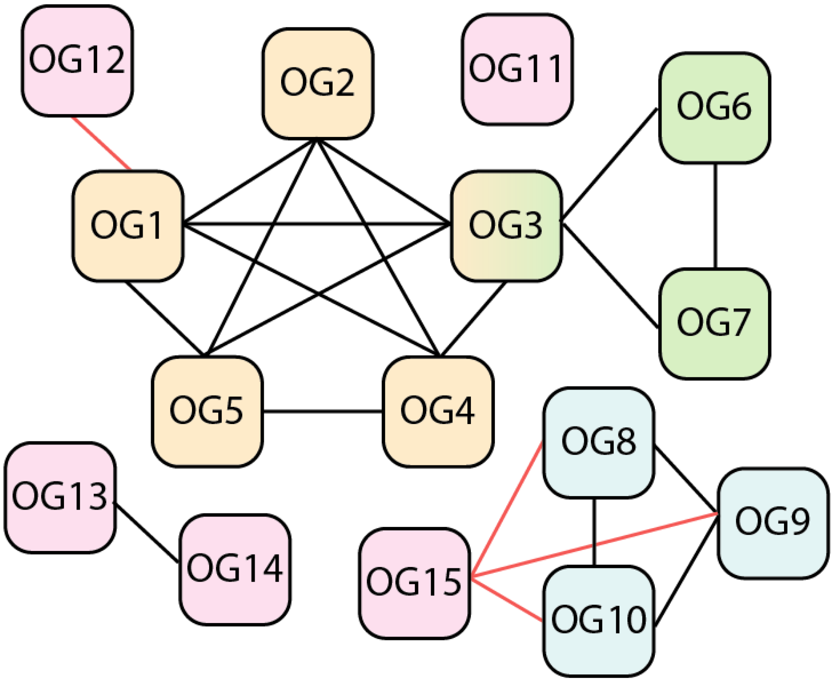
16 - Figure 3.4 - Rearrangements in gene neighbourhoods

The distribution of differences between conservation of the most conserved ortholog group in a neighbourhood (total proportion of conservation in a neighbourhood relative to the seed gene) and its highest positional conservation (proportion of conservation at a certain gene distance from the centered gene relative to the seed gene). Only neighbourhoods with at least one ortholog group with a neighbourhood conservation of 20% were included.

3.2 Identification of conserved gene clusters

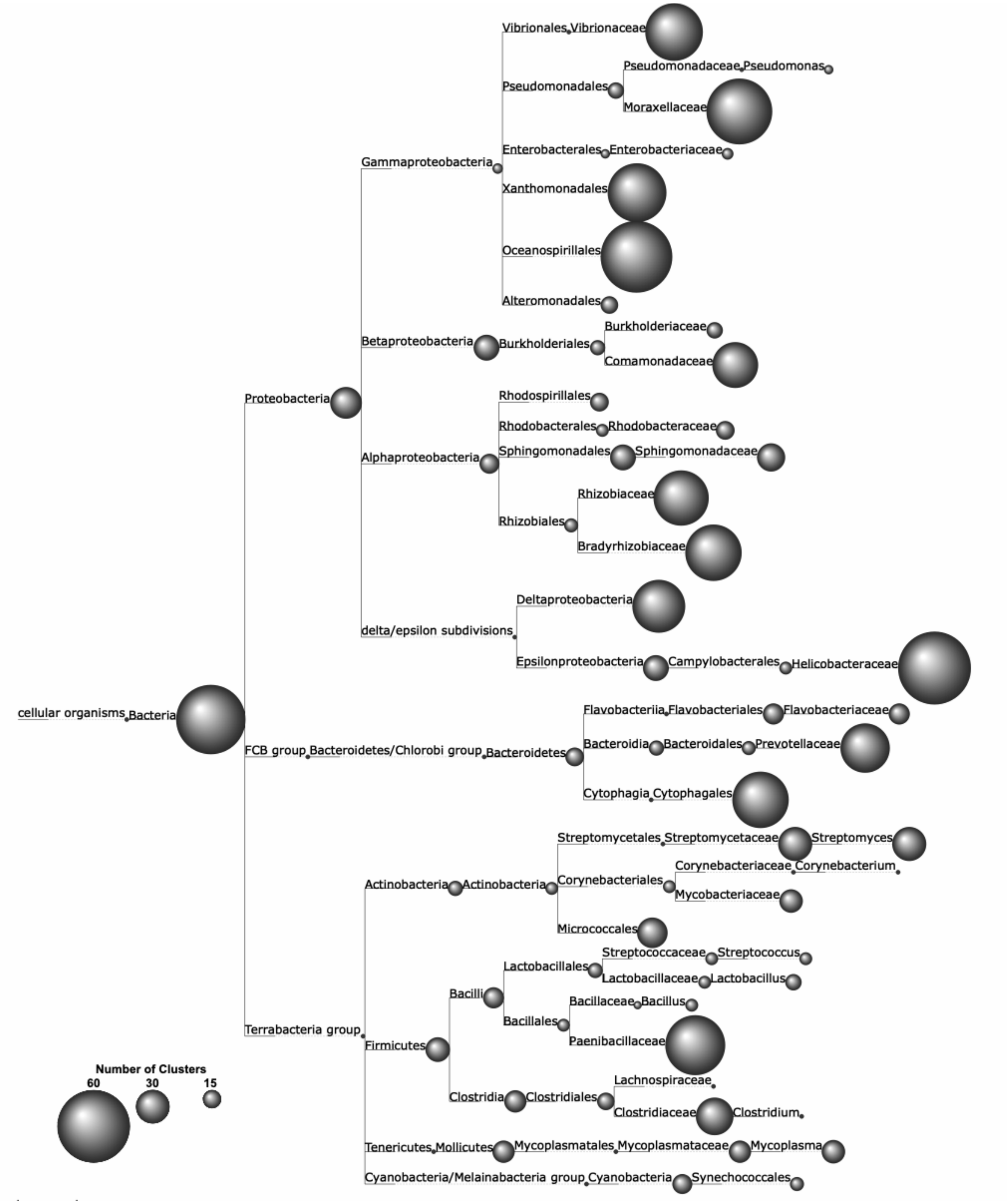
Method of identification

The gene level resolution of our dataset provided us the ability to generate genome context data for thousands of genomes rapidly, and then analyse these neighbourhoods at the gene order level. However, with a large amount of data comes complexity. Therefore, in order for us to better study gene clusters we devised a way to consolidate our dataset into gene clusters by grouping closely conserved genes together. Genes were networked and connected by their reciprocal co-conservation values (Figure 3.5). Connections which were below the significance threshold as defined in Chapter 2 were removed (Figure 2.3). All genes that were then still networked to two or more other genes were taken together as a gene cluster. To prevent very large clusters from being classified due to the presence of one gene in two different clusters, all bridge genes (those with neighbours whom did not network with each other) were duplicated, with each clone taking a separate group of neighbours, forming two separate clusters (Figure 3.5). Using this strategy, we could repeat this process at all taxonomic levels allowing us to determine when gene clusters first arose and how they changed over evolutionary time. By implementing this technique, we identified 1383 gene cluster families over 35 different taxonomic nodes (Figure 3.6 and Table 8.7). In total this consisted of 4827 gene clusters when including the many variations of a gene cluster family seen at the different taxonomic levels. We found that neither the number of gene neighbourhoods nor the number of clustered genes were overrepresented in a particular bacterial taxon however in more recent taxonomic divisions, the number of detected clusters increases, most likely due to reduced evolutionary time for genomic shuffling to occur and the increased influence of HGT (Figure 3.6).



17 - Figure 3.5 - Schematic of the clustering algorithm

A representation of how our clustering algorithm functions. Each square represents a different ortholog group. The lines between them represent the neighbourhood conservation between the genes, and are coloured red when this falls below the threshold. The ortholog groups are coloured based on the clusters they are assigned, except for red which identifies ortholog groups which are classified as not in a gene cluster. Multiple colours represent the multiple different clusters an ortholog group can belong to.


Vibrionales
Vibrionaceae
Pseudomonadaceae
Pseudomonas
Pseudomonadales
Moraxellaceae
Enterobacterales
Enterobacteriaceae
Gammaproteobacteria
Xanthomonadales
Oceanospirillales
Alteromonadales
Burkholderiaceae
Betaproteobacteria
Burkholderiales
Comamonadaceae
Proteobacteria
Rhodospirillales
Rhodobacterales
Rhodobacteraceae
Sphingomonadales
Sphingomonadaceae
Alphaproteobacteria
Rhizobiaceae
Rhizobiales
Bradyrhizobiaceae
Deltaproteobacteria
delta/epsilon subdivisions
Epsilonproteobacteria
Campylobacterales
Helicobacteraceae
cellular organisms
Bacteria
Flavobacteriia
Flavobacteriales
Flavobacteriaceae
Bacteroidia
Bacteroidales
Prevotellaceae
FCB group
Bacteroidetes/Chlorobi group
Bacteroidetes
Cytophagia
Cytophagales
Streptomycetales
Streptomycetaceae
Streptomyces
Corynebacteriaceae
Corynebacterium
Actinobacteria
Actinobacteria
Corynebacteriales
Mycobacteriaceae
Micrococcales
Streptococcaceae
Streptococcus
Lactobacillales
Lactobacillaceae
Lactobacillus
Bacilli
Bacillaceae
Bacillus
Bacillales
Terrabacteria group
Paenibacillaceae
Firmicutes
Number of Clusters
60
30
15
Lachnospiraceae
Clostridia
Clostridiales
Clostridiaceae
Clostridium
Tenericutes
Mollicutes
Mycoplasmatales
Mycoplasmataceae
Mycoplasma
Cyanobacteria/Melainabacteria group
Cyanobacteria
Synechococcales

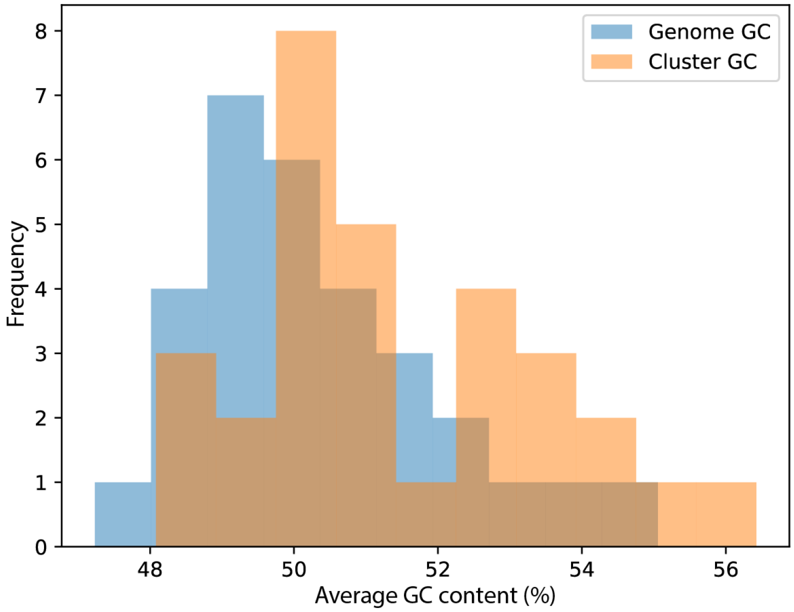
18 - Figure 3.6 - Identification of gene clusters in bacteria

Gene neighbourhoods were clustered together based on common conservation at each taxonomic division. Clusters that contained 50% or more genes from a cluster identified at an ancestor clade were considered a descendant cluster and not unique. The size of each circle represents the number of unique clusters (the oldest member of each gene cluster family) identified at each taxonomic level (see the legend in the bottom left corner).

The role of HGT in gene clusters

As we detailed in the introduction, horizontal gene transfer is a major factor in bacterial chromosome evolution. Unlike vertical transmission where gene clusters are transmitted to descendants and therefore are taxonomically isolated, horizontal gene transfer permits the spread of genes clusters independent of phylogenetic relatedness. To analyse and compare these two modes of gene cluster transmission we used two normalization strategies for conservation during the clustering process. Our initial strategy normalizes the frequency two genes co-localize by the number of genomes belonging to the taxonomic distribution. This results only in clusters which are prevalent in the majority of the clade being identified. As this precludes horizontal transmission, we also used a strategy independent of taxonomic definitions where conservation was normalized by the number of genomes each ortholog group appeared in (the same strategy used to calculate neighbourhood and positional conservation from histograms). This strategy does not bias against genes which are dispersed across the Bacteria kingdom. Therefore, gene clusters which uniquely appear using this normalization strategy are mostly distributed via HGT as opposed to vertical transmission. A limitation of this strategy is that neighbourhoods which are highly taxonomically restricted would both not have appeared in our previous analysis and would be ranked very highly even if the genes were not involved in HGT. Therefore, we enforced a minimum required genome distribution of 50. Using this strategy, we identified 59 clusters which were not detected using the first method (Table 8.8). As this number is significantly lower than the taxonomic bound strategy, this suggests that over large evolutionary time scales HGT is not relevant in conserved gene clusters. Analysing the HGT clusters, we saw that they were smaller on average than those found from vertical transmission with a maximum size of 5 genes, and only 11 being larger than three genes (Table 8.8). There are often size limitations with the number of nucleotides (and therefore genes) that can be transferred via most HGT methods which could explain the smaller size of these clusters. The majority of HGT associated gene clusters were poorly characterized but often had annotated membrane domains or were thought to be secreted (Table 8.8). A few noticeable clusters included a cluster of pilus assembly genes which would be required for HGT method of conjugation. There were also several clusters containing ABC transporters, which often have roles in virulence and antibiotic resistance which are genes they are often implicated in HGT (Derbyshire and Gray 2014).

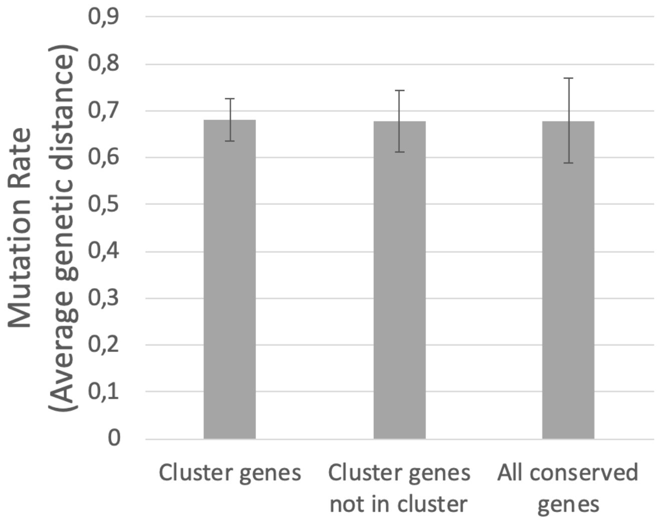
Genes which are horizontally transferred have been observed to have reduced GC content (Daubin, Lerat, and Perrière 2003). It is thought this bias may occur due to a few factors which favour the transmission of sequences rich in AT. These include: the majority of bacterial restriction enzymes are biased towards GC sequences resulting in high GC genes being less likely to be incorporated into the genome as free DNA, and both phages and insertion sequences, two other vectors of HGT, favoring sequences low in GC content (Rocha and Danchin 2002). In order to observe to test how likely horizontal gene transfer may have played a role in forming and spread gene clusters we measured the GC content of the clustered regions compared to the host genome. Clusters were shown on average to have a higher GC content than the whole genome by 2% (Figure 3.7). This relationship held in clusters from more recent taxonomic divisions as well. This thereby suggests both an unlikely origin via HGT for many clusters and that the clusters themselves are less likely than other parts of the genome to be transferred. This implicates vertical transmission as the major defining factor for gene cluster organization.



19 - Figure 3.7 - GC content bias of clusters

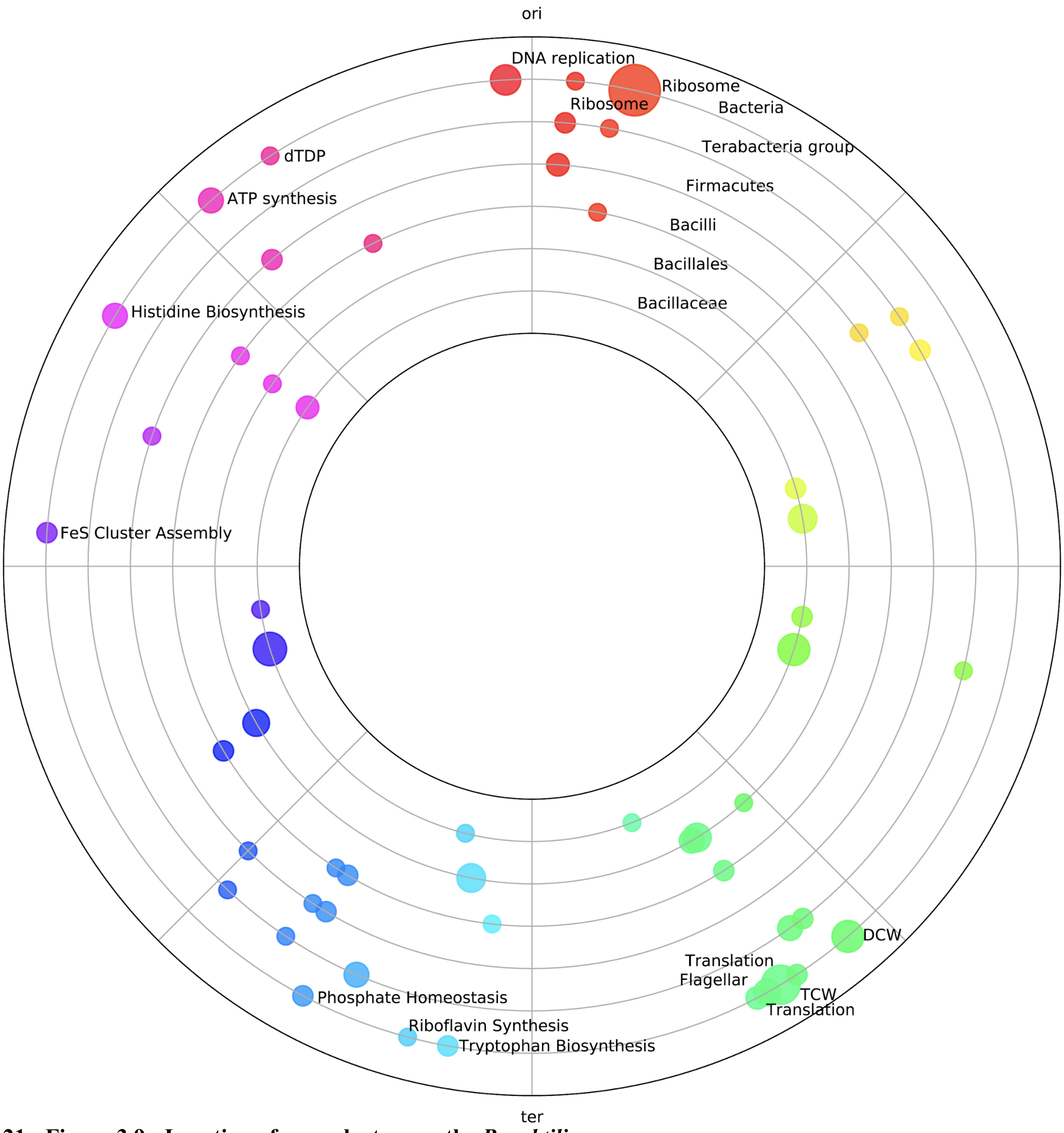
The GC content of clusters was calculated as the average GC content of the genomic cluster spanning region in all genomes where the cluster had a minimum of four of the genes co-localized. Then the average GC content of this subset of genomes was measured and compared.

In bacteria mutations have been shown to be biased towards AT (Hershberg and Petrov 2010), suggesting that areas rich in GC content undergo less maintained mutations. As we see an increased rate of GC occurrence in gene clusters, the possibility that clusters form in areas protected from mutagenesis thereby resulting in clusters of essential processes. We measured the mutation rate of genes found in conserved neighbourhoods compared to equally conserved genes not found in neighbourhoods and found that neighbourhood genes did not undergo significantly more or less mutations than non-clustered genes of similar ubiquity (Figure 3.8). Additionally, we saw no difference between the mutation rates of clustered genes compared to the same genes in genomes where they are no longer clustered (Figure 3.8).



20 - Figure 3.8 - Mutation rate of clustered genes

The average sequence distance (as measured by Clustalw) was measured, comparing all sequences of genes from clusters identified at the bacterial level, that were still found with 50% of their cluster members in extant genomes. This was repeated for cluster gene sequences which were no longer associated with the cluster and for all highly conserved genes >2000 genomes. Error bars represent the standard deviation of the samples.



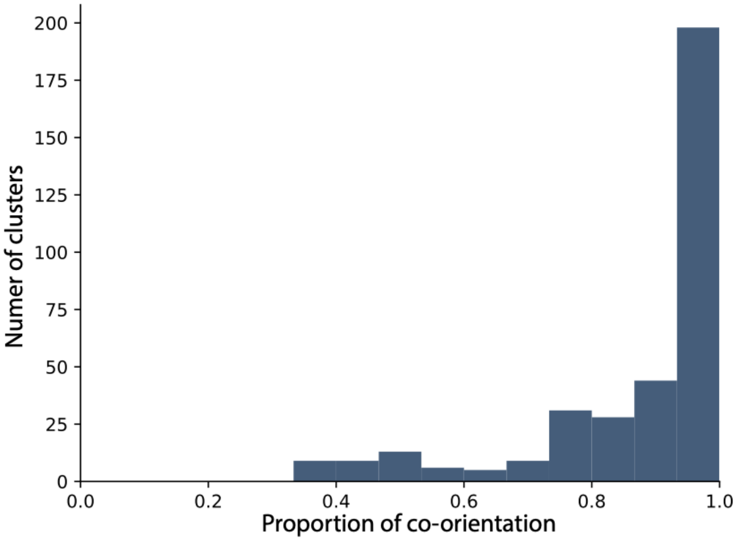
21 - Figure 3.9 - Location of gene clusters on the *B. subtilis* genome

The genomic location of all gene cluster families presents on the *B. subtilis W168* genome. The y-axis represents at which taxonomic division the cluster was first detected. Clusters detected at the Bacteria level are labeled with a custom annotation based on the function of genes in the cluster. The size of the circle represents the number of genes in the cluster. The colour is another indicator of the position of the chromosome.

3.3 Analysis of gene cluster conservation

Co-orientation is highly conserved in gene clusters

In Chapter 3.1 we noted that co-orientation rapidly degrades with gene distance (Figure 3.1). However, we also saw high correlation between orientation and conservation in neighbourhoods (Figure 3.2). To test which of these two forces is stronger on conserved gene clusters, which are both long and well conserved, we measured the average co-orientation that occurs in our identified clusters. We found that indeed co-orientation was highly conserved in gene clusters irrespective of the gene length of the cluster (Figure 3.10). This further strengthens the correlation we saw between gene neighbourhood conservation and orientation and suggests that co-orientation is an important factor in the maintenance of gene clusters.

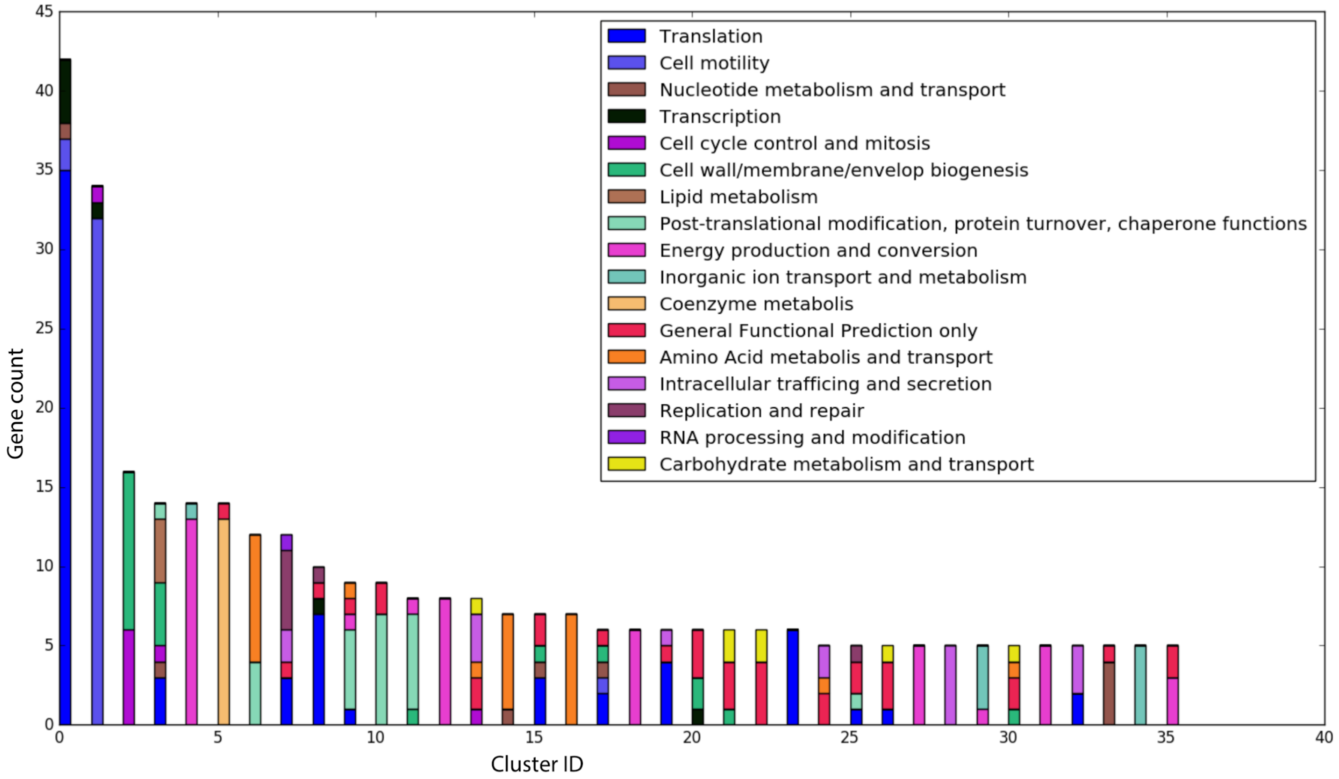


22 - Figure 3.10 - Conservation of orientation within gene clusters

A histogram showing the number of gene clusters with specific proportions of co-orientation of all involved genes. Only the oldest member of each gene cluster family was used. In cases where a gene from the cluster was not present on the genome, and therefore had no orientation, this contributed a 0 to the average and explains how values under 0.5 are possible.

Function conservation within clusters

One of our goals is to understand the overarching forces that drive gene cluster formation, tangential to this, is if particular cellular functions benefit more from clustering their individual genes with others. We annotated the clusters identified in Bacteria using a manually curated list of gene ontology terms (GO terms) (The Gene Ontology Consortium 2019), designed to capture ubiquitous and clear defined biological processes in the cell. We found that several clusters contain genes with homologous functions, which was most apparent in the larger clusters whereas smaller clusters had genes with a diverse range of functions (Figure 3.11). It has been previously reported that clusters are functionally homogeneous (Wolf, Rogozin, Kondrashov, et al. 2001; R. Overbeek et al. 1999). We presume, that the small but diverse clusters in our dataset were detected due to the higher sensitivity of our detection method and were not identified in previous studies.



23 - Figure 3.11 - Functional roles of cluster genes

For gene clusters identified at the bacterial level with five or more genes, we classified each of their gene members using a predefined gene ontology list (see legend for GO terms). The colour represents which annotation the gene was given (see the legend in the top right corner). Clusters are sorted by number of genes

To get a quantitative understanding of gene function in clusters we performed enrichment analysis with the genes found in clusters of *B. subtilis* using the genome as a background (Figure 3.8). Using the web-tool DAVID (D. W. Huang, Sherman, and Lempicki 2009), we looked for enrichments in GO terms, uniprot keywords, KEGG pathways and Interpro domains. We found nine groups of annotations that appeared to be enriched in clustered genes including: ribosome/translation genes, ATP synthases, flagella genes, pyrimidine/arginine biosynthesis genes, sigma factors, cell shape associated genes, iron binding, and DNA repair. (Table 3.1). 16 significant terms (<0.05 P-value Bejamini Hochburg correction) were not clustered. This was mostly unchanged when restricting the clustered genes to only those detected at the bacteria level. The majority of these associates align with well known gene clusters such as the ribosome suber-cluster and the ATP synthase cluster.

Enriched Gene Annotations in Clustering Genes

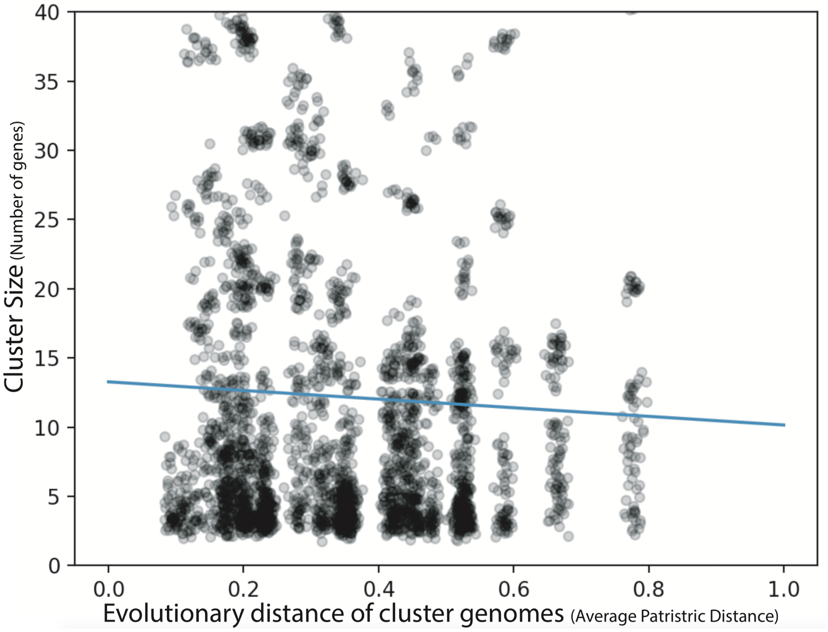
Category	Term	Count	Pop Hits	Fold Enrichment	PValue	Benjamini
Annotation Cluster 1	Enrichment Score: 18.03833735983574					
GOTERM_MF_DIRECT	GO:0003735~structural constituent of ribosome	49	55	3.72	9.43E-24	3.06E-21
GOTERM_BP_DIRECT	GO:0006412~translation	52	62	2.98	1.45E-19	3.00E-17
GOTERM_MF_DIRECT	GO:0019843~rRNA binding	34	38	3.73	1.96E-16	3.61E-14
GOTERM_CC_DIRECT	GO:0005840~ribosome	39	49	3.22	2.62E-15	1.17E-13
Annotation Cluster 2	Enrichment Score: 2.0645164271685235					
GOTERM_MF_DIRECT	GO:0046933~proton-transporting ATP synthase activity, rotational mechanism	7	7	4.17	1.02E-03	5.38E-02
GOTERM_BP_DIRECT	GO:0015986~ATP synthesis coupled proton transport	7	7	3.55	2.57E-03	2.33E-01
GOTERM_CC_DIRECT	GO:0045261~proton-transporting ATP synthase complex, catalytic core F(1)	5	5	4.05	1.46E-02	7.78E-02
GOTERM_MF_DIRECT	GO:0046961~proton-transporting ATPase activity, rotational mechanism	3	3	4.17	1.44E-01	9.49E-01
Annotation Cluster 3	Enrichment Score: 1.3463031253112354					
GOTERM_MF_DIRECT	GO:0003774~motor activity	5	5	4.17	1.31E-02	3.01E-01
GOTERM_CC_DIRECT	GO:0009425~bacterial-type flagellum basal body	6	8	3.03	2.53E-02	1.18E-01
GOTERM_CC_DIRECT	GO:0031514~motile cilium	16	38	1.70	2.89E-02	1.21E-01
GOTERM_BP_DIRECT	GO:0071973~bacterial-type flagellum-dependent cell motility	9	16	2.00	5.16E-02	7.46E-01
GOTERM_BP_DIRECT	GO:0006935~chemotaxis	5	11	1.61	3.75E-01	9.83E-01
Annotation Cluster 4	Enrichment Score: 1.2374018318599636					
GOTERM_BP_DIRECT	GO:0044205~'de novo' UMP biosynthetic process	8	10	2.84	7.00E-03	2.15E-01
GOTERM_MF_DIRECT	GO:0004088~carbamoyl-phosphate synthase (glutamine-hydrolyzing) activity	4	5	3.34	9.22E-02	8.77E-01
GOTERM_BP_DIRECT	GO:0006526~arginine biosynthetic process	5	10	1.77	3.01E-01	9.75E-01
Annotation Cluster 5	Enrichment Score: 1.1058545075576294					
GOTERM_MF_DIRECT	GO:0016987~sigma factor activity	11	20	2.30	9.75E-03	2.73E-01
GOTERM_BP_DIRECT	GO:0006352~DNA-templated transcription, initiation	10	19	1.87	5.70E-02	7.03E-01
GOTERM_MF_DIRECT	GO:0003700~transcription factor activity, sequence-specific DNA binding	23	110	0.87	8.66E-01	1.00E+00
Annotation Cluster 6	Enrichment Score: 0.9767810932765434					
GOTERM_MF_DIRECT	GO:0004129~cytochrome-c oxidase activity	6	7	3.58	1.04E-02	2.67E-01
GOTERM_CC_DIRECT	GO:0070469~respiratory chain	3	4	3.03	2.55E-01	6.61E-01
GOTERM_MF_DIRECT	GO:0005507~copper ion binding	3	6	2.09	4.40E-01	9.98E-01
Annotation Cluster 7	Enrichment Score: 0.9651972102490863					
GOTERM_BP_DIRECT	GO:0051301~cell division	15	33	1.61	5.29E-02	7.13E-01
GOTERM_BP_DIRECT	GO:0007049~cell cycle	11	22	1.77	6.09E-02	6.94E-01
GOTERM_BP_DIRECT	GO:0008360~regulation of cell shape	9	18	1.77	1.02E-01	8.20E-01
GOTERM_BP_DIRECT	GO:0009252~peptidoglycan biosynthetic process	7	13	1.91	1.28E-01	8.68E-01
GOTERM_BP_DIRECT	GO:0071555~cell wall organization	6	14	1.52	3.54E-01	9.81E-01
Annotation Cluster 8	Enrichment Score: 0.24307645512331666					
GOTERM_MF_DIRECT	GO:0005506~iron ion binding	8	21	1.59	2.18E-01	9.82E-01
GOTERM_MF_DIRECT	GO:0016705~oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen	3	9	1.39	6.73E-01	1.00E+00
GOTERM_MF_DIRECT	GO:0004497~monooxygenase activity	3	12	1.04	8.21E-01	1.00E+00
GOTERM_MF_DIRECT	GO:0020037~heme binding	5	25	0.83	8.83E-01	1.00E+00
Annotation Cluster 9	Enrichment Score: 0.11282483408671384					
GOTERM_BP_DIRECT	GO:0009432~SOS response	4	10	1.42	5.66E-01	9.97E-01
GOTERM_BP_DIRECT	GO:0006281~DNA repair	7	28	0.89	8.43E-01	1.00E+00
GOTERM_BP_DIRECT	GO:0006310~DNA recombination	5	26	0.68	9.61E-01	1.00E+00

2 - Table 3.1 - Enriched annotations in *B. subtilis* clustered genes

A list of the grouped GO term annotations which were enriched amongst clustered genes compared to all the genes on the *B. subtilis* genome. Annotation clusters are sorted from most enriched to least. Count represents the number of times the annotation was in the clustered genes subset. Pop Hits, is the number of times the annotation was found in all genes in genome.

Growth dynamics of conserved gene clusters

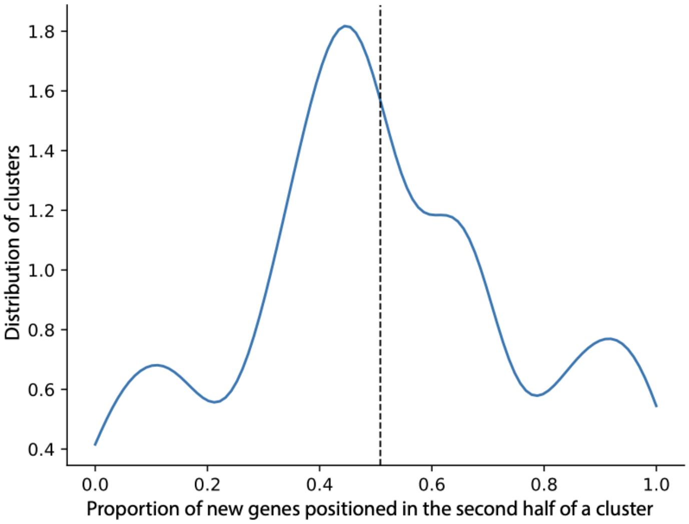
As a definition, gene clusters must be resistant to genomic rearrangements which would separate the genes. We asked if this resistance extends to nearby genes as the number of possible (those that result in fit offspring) rearrangement events that could move them would decrease. If so, under a neutral model of genomic rearrangement (Darling, Miklós, and Ragan 2008) this would be evidenced by an expansion in the size of gene clusters over evolutionary time. Therefore, we tracked the size of gene clusters down each gene cluster family lineage. Gene cluster families and their lineages were defined as follows: for every identified cluster, if another cluster at a higher taxonomic rank shared more than 50% of its gene content with it, that cluster would be considered the ancestor of this cluster. We work under the parsimonious assumption that if two species share a gene cluster, that this was likely derived from vertical inheritance from a shared common ancestor as opposed to being formed in parallel. Given the stringency of our cluster classification method the effect of taxonomically dispersed HGT clusters should be minimal. Naturally the strength of neighbourhood conservation surrounding gene clusters increases as we restricted the genome subset. As our clustering uses the threshold for expected conservation at a specific genetic diversity that we calculated in Chapter 2, we control for the conservation increase. Therefore, any addition to a gene cluster must be conserved significantly higher than would be expected by chance. We saw however that on average there was neither a significant increase or decrease in gene cluster size in more closely related genome samples (Figure 3.12 - blue line). On average we see ~2 genes added to a gene cluster over the span of the entire bacterial lineage (Figure 3.12). Of the lineages we were able to track, clusters were seen to expand and shrink in 46.6% and 27% of cases respectively. The remaining 25.9%, the cluster size remained unchanged. Therefore, this suggests clusters do not act as anchors for further clustering of genes and are more likely to be generated spontaneously (Figure 3.12). Simultaneously, this suggests that clusters are not slowly eroded through evolutionary history but are quickly dispersed all together.



24 - Figure 3.12 – Size dynamics of gene clusters over evolutionary time

The cluster size of each gene cluster family member is plotted against the genetic diversity (the average patristic distance between all genomes of the taxa) of the taxonomic division it was clustered on. A linear regression was fitted to the data (blue). A slight jitter has been applied to each point for clarity purposes.

The coexpression of genes as operons is thought to be a driving force in the maintenance of gene clusters (Fani, Brilli, and Liò 2005). We have also observed strong biases in the correlation of orientation both within gene neighbourhoods and the identified gene clusters (Figure 3.2 and 3.10). Therefore, we reasoned, if genes do indeed cluster in order to be regulated under common regulatory factors, new genes should appear primarily downstream of the cluster to not disrupt regulation of other genes by e.g. the promoter. Therefore, we measured the number of new genes added to growing clusters which appeared in the downstream half of the gene cluster (orientated relative to the majority of genes in the cluster). We found that there appeared to be no preference to either side in the context of new genes being added the clusters (Figure 3.13). This places into doubt the role of co-transcription in cluster formation.

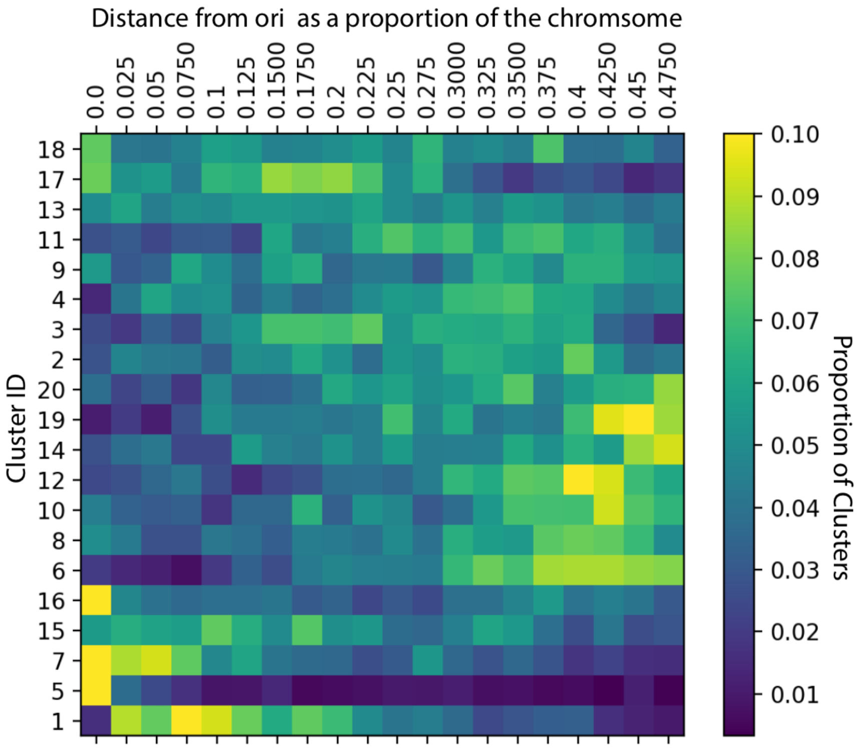


25 - Figure 3.13 - Directional Growth of clusters

The lineage of each gene cluster family was tracked down the evolutionary tree. We performed a kernel density estimation on the proportion of new members of growing gene clusters were positioned either in the second half or downstream of the cluster. Clusters were considered descendants if they contained 50% of the genes from an ancestor cluster.

Gene Clustering and genome organization

Two well known gene clusters (the DNA replication and the ribosome cluster) are known to be localized close to the origin of replication (ori) (Couturier and Rocha 2006). It is thought that the conserved location of these clusters may be due to the increased gene dosage effects they undergo during replication (Couturier and Rocha 2006). It has also been shown that the localization of genes on the genome can affect the subsequent localization of the genome (Ginez, Osorio, and Poggio 2014). Therefore, we asked if genome organization was an important intrinsic property to the conservation of clusters. By taking the median distance of clusters from the ori on each genome we were able to define four different conservation behaviors: ori localized (>50% of clusters found in the first quintile), terminator localized (>50% of clusters found in the last quintile), non-polar localization (>50% of clusters found in the second to fourth quintile), and no location conservation. The majority of localized clusters we found to be either ori or terminus proximal (46.6%) with only one cluster localizing between the ori and terminus regions and the remainder not localizing (Figure 3.14). We see that ori proximal clusters are enriched in translation machinery but with also the notable inclusions of secretion/transport systems and the ATP synthase cluster. As mentioned previously, ori proximal genes undergo a gene dosage increase which is magnified in conditions of fast growth where multifork replication occurs. During these conditions the bacterial cells have an increased demand for translation machinery, ribosomes and ATP (Couturier and Rocha 2006). Conversely, clusters which were located at the terminal end of the chromosome did not have a unifying function however they were just as abundant as those seen near the ori. In addition, there is the presence of a smaller ribosome cluster and a translation cluster at the terminal end which would receive much lower transcriptional activity due to a lesser effect from gene dosage. The presence of phage genes near the terminus can be explained as horizontal gene transfer since pro-phage activity occurs more frequently in genomic areas closer to the terminus (Oliveira et al. 2017) We found that very few clusters were specifically localized between either of the genome poles however there were many clusters which displayed no localization at all.

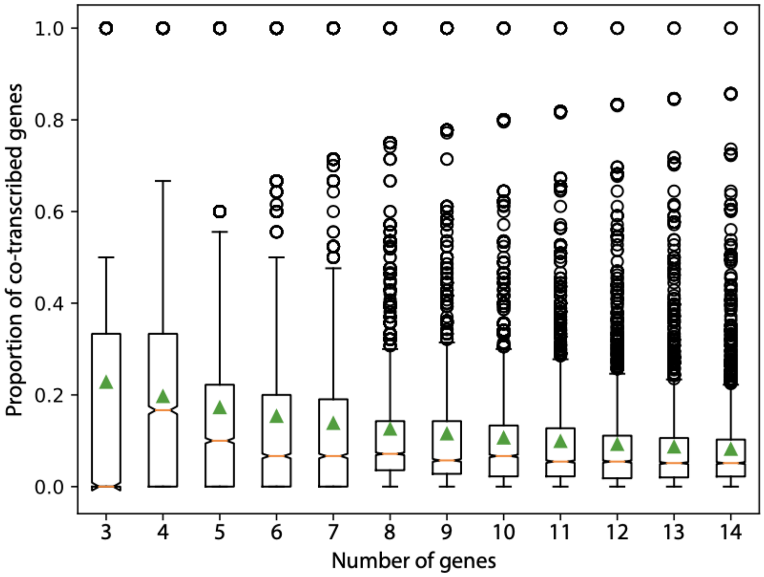


26 - Figure 3.14 - Genomic location of gene clusters

For the gene clusters identified at the bacterial level, their position in their respective genomes relative to the oriC were compiled. x-axis labels represent bins of distance (5% of the distance between the origin of replication and the terminus). Shading of the bins is based on the proportion of genomes with the cluster which had the cluster localized in this area. Rows were clustered based on similarity. Clusters matching the Cluster ID can be found in Table 8.7. Cluster IDs are sorted by size of the cluster in descending order.

Selective pressures of gene clusters

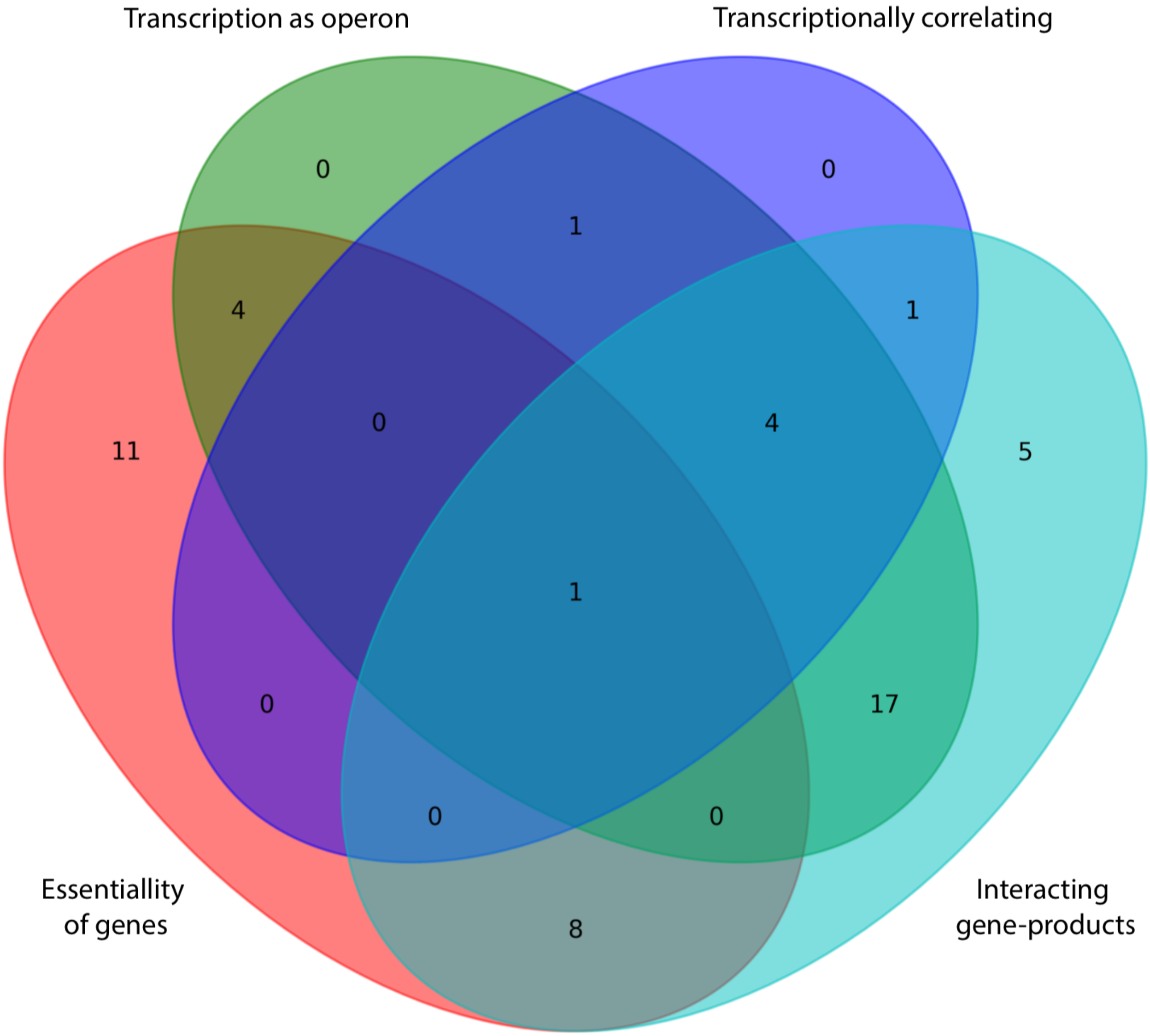
There are several hypotheses as to which selective forces act on bacterial genomes to preserve gene neighbourhoods including: essentiality of the genes, protein-protein interactions of gene-products of the cluster and co-transcription on polycistronic transcripts. Given our large number of clusters scattered over the evolutionary history of an organism we now have the possibility to see how or if these selective forces come into play. Firstly, we measured the average proportion of a group of genes, from the *B. subtilis* genome, that are essential, interacted with, or is transcribed in the same transcript, as another gene from the cluster. We also measured the average Pearson correlation score of all pairwise combinations of cluster genes as a measure of transcriptional correlation using publicly available RNAseq datasets for *B. subtilis* (for conditions see Table 8.9). These classifications were assigned to the genes using the DEG10 database for essentiality (Luo et al. 2014), the DOOR database for operons (Mao et al. 2009) and STRING database for protein-protein interactions (Szklarczyk et al. 2019). We used two methods to determine the expected proportion, taking the average of both gene groups derived from a sliding window along the genome, and 100,000 randomized gene groups (An example can be seen in Figure 3.15). As the size of the gene groups could be a confounding factor, we repeated this process for several different sizes so that appropriate comparisons could be made (Figure 3.15). Unlike essentiality and protein-protein interactions, operons are required to have their genes next to each other and therefore measuring operon proportionality using randomized gene groups was not measured in this case. Using all clusters which were still maintained in the extant *B. subtilis* species we compared to what is seen across the genome and they were considered significant if the proportion in the cluster was found to be higher than one standard deviation above the mean. The mean and standard deviation were taken from the method which provided the higher values therefore to increase stringency of our method. The thresholds (mean + standard deviation) for percentage of: essentiality, protein-protein interactions, co-transcribed genes, and average Pearson correlation score for clusters of size 3 are respectively: 5.34% + 13.4% (18.74%), 14.4% + 23.3% (37.7%), 22.8% + 34.3% (57.1%), 61.1% + 29.7% (90.8%).



27 - Figure 3.15 - Calculation of expected operons found in gene clusters using a sliding window approach

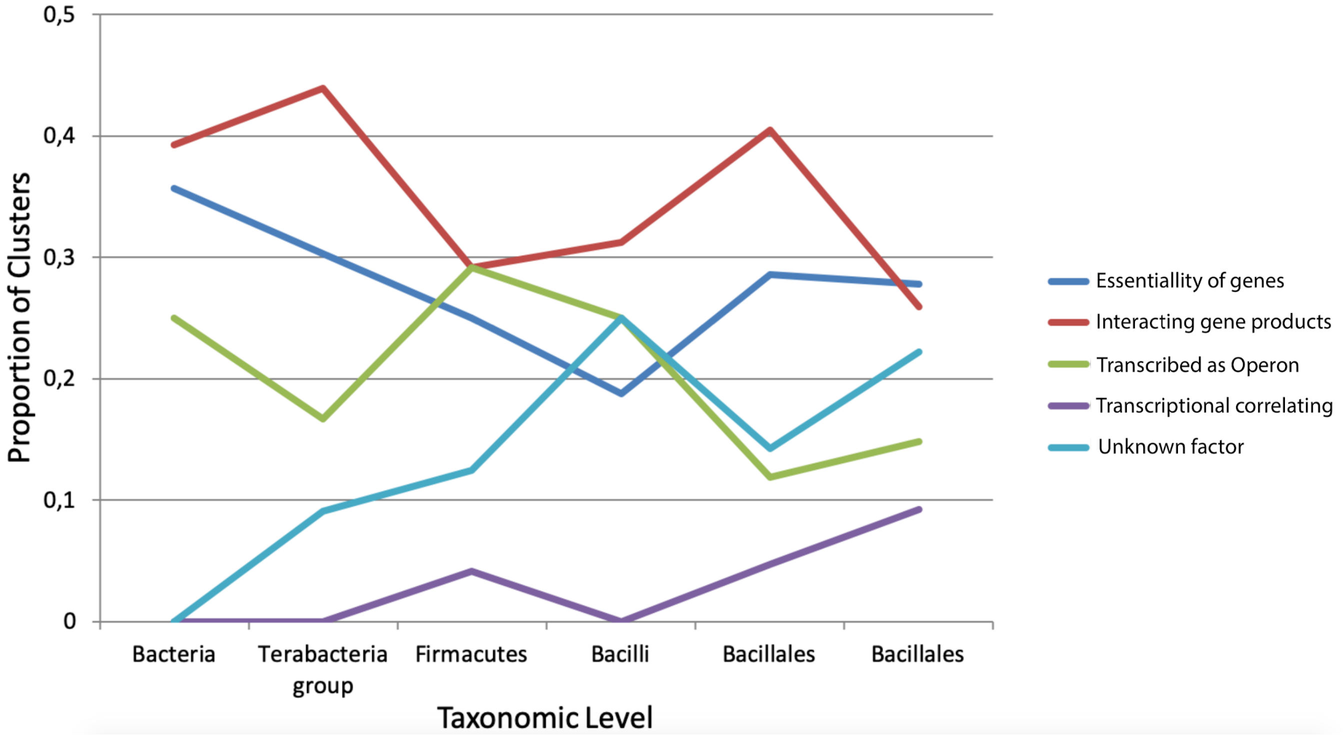
The proportion of genes which were found to be transcribed on the same transcripts of gene groups from a sliding window on the *B. subtilis* genome. Operon status was defined by the DOOR database. Process was repeated using different sliding window sizes reflecting the number of genes (x-axis). Green triangle represents the mean and the yellow line represents the median. The whiskers extend to highest datum within 1.5 interquartile range of the upper quartile and vice versa. Flyers (values greater or less than the whiskers) are displayed as circles.

We found that essentialness of genes, protein-protein interactions and operon transcription was enriched in neighbourhoods in 42.1%, 63.2% and 47.3% of clusters, however transcriptional co-regulation, was enriched in only 13.5% of clusters (Figure 3.16). That operons were enriched but not transcriptional co-regulated may appear contradictory however one gene cluster could contain multiple operons which are differently regulated. Furthermore, the measured co-regulation of *B. subtilis* genes was very high compared to the average number of operons, therefore it is harder to be significantly corregulating. Finally, as co-regulation is possible through other means independent of genome context such as in regulons this is unsurprising. Interestingly, it was observed that whilst operons were more enriched in neighbourhoods it was never the only selective force representing a cluster unlike that which was seen for interaction and essentialness (Figure 3.16). This suggests that simply having genes in an operon structure or co-regulated is not a strong enough evolutionary force to keep gene neighbourhoods and may be a consequence of having genes together over a long evolutionary time. To get a better understanding of how theses selective forces shaped the evolution of gene clusters, we delineated the contributions based on the taxonomic level for which we first detected the clusters. Transcriptional correlation has little impact in older clusters and only appears as a factor in more recently observed gene clusters whereas the impact of the other selective forces remained relatively stable (Figure 3.16). We found that many recently occurring gene clusters could not be explained by any of the factors mentioned above. Interestingly all of theses clusters were localized either by the origin or terminus of the genome further implicating these regions as hotspots for cluster formation.



28 - Figure 3.16 - Evidence of selective pressures in gene clusters

All extant gene clusters found in *B. subtilis* were measured for the abundance in essential proteins, proteins which co-interacted, genes which are coregulated and genes that were transcribed together on polycistronic transcripts. These were compared to the expected genome frequency as calculated both as a sliding window and random selection from the genome at multiple window sizes. Clusters were considered abundant in the selective factor if the proportion was one standard deviation higher than the genome average. Essentiality, interactions and operon status in *B.subtilis* were defined by the DEG10 database for essentiality (Luo et al. 2014), the DOOR database for operons (Mao et al. 2009) and STRING database for protein-protein interactions (Szklarczyk et al. 2019).



29 - Figure 3.17 - Relevance in selective pressures over time

All extant gene clusters found in *B. subtilis* were measured for the abundance in essential proteins, proteins which co-interacted, genes which are corregulated and genes that were transcribed together on polycistronic transcripts. These were compared to the expected genome frequency as calculated both as a sliding window and random selection from the genome at multiple window sizes. Clusters were considered abundant in the selective factor if the proportion was one standard deviation higher than the genome average. Essentiality, interactions and operon status in *B.subtilis* were defined by the DEG10 database for essentiality (Luo et al. 2014), the DOOR database for operons (Mao et al. 2009) and STRING database for protein-protein interactions (Szklarczyk et al. 2019). Clusters were only assigned "unknown factor" when they were not enriched in any of the aforementioned factors. The taxonomic levels on the x-axis are sorted from broadest to narrowest taxa containing *B. subtilis*.

3.5 Summary

With the resources we generated from GenCoDB we had access to gene neighbourhood data at a resolution previously not possible. This facilitated the comparisons of genomic contexts of genes from 4,036,537 ortholog groups encompassing 5487 genomes over 89 different taxonomic levels, which together summed over 1.9 million neighbourhoods. Despite the expectation that neighbourhood conservation should be anticorrelated with how conserved the genes are, we found that several genes which were conserved across the bacterial kingdom had highly conserved gene neighbourhoods. This also highlighted several genes which were less conserved but had higher neighbourhood conservation than average which would normally not be detectable with fewer genomes. In looking at neighbouring genes we saw that conservation and co-orientation both correlate and rapidly degrade as a function of distance resulting in clusters greater than three genes being rare. Conversely, we found that once co-localization was established, short-range rearrangements within the neighbourhood are very rare. We developed a method to define gene clusters using our gene neighbourhood data finding 1383 gene cluster families. A complementary analysis identified 59 clusters that were likely distristributed by horizontal gene transfer despite the finding that the majority of our identified gene clusters were anti-correlated with characteristics associated with HGT. Whilst orientation rapidly degrades as neighbourhoods get larger, irregardless of the size of the gene clusters we found co-orientation was highly present. However, this does not lead to a bias in cluster growth. On average, we saw cluster growth was not the standard suggesting that conserved neighbourhoods do not encourage the conservation of more genes around it. Furthermore, we observed that gene clusters were highly enriched in the origin of replication and terminal parts of the chromosome. We tested the prevalence of the hypothesized selective pressures of gene clusters: essentiality of the genes, protein-protein interactions within genes in the cluster and co-transcription on polycistronic transcripts, and indeed found an enrichment of all three. Strikingly, whilst many clusters had a mix of selective pressures, or contained interacting or essential genes, very few were solely under the selective pressure created by operon level transcription. This suggests that this is a weak evolutionary force and may not be enough independently to maintain gene clusters on a fluid genome.

4. Perturbation of a conserved translation and cell envelope synthesis associated gene cluster

In this chapter we will apply GenCoDB, the quantitative bacterial genome context analysis web-tool, to the task of identifying a conserved gene cluster connecting translation and cell envelope synthesis. Using the findings on how evolution acts on gene clusters in bacteria that we learned about in chapter 3, we will bioinformatically and genetically interrogate the gene cluster in an attempt to decipher the fitness benefits from the co-localization of these genes in the model organism *Bacillus subtilis*. We will highlight that several key theories of gene neighbourhood conservation do not hold for this cluster including that polycistronic transcripts do not play a key regulatory role in correlating gene expression. Finally, we will present our own model on how this cluster may act through pyrimidine/arginine metabolism and consequently central carbon metabolism to synchronize essential processes in the cell.

4.1 Identification of genomic-linked volume/surface area mediators

Earlier we introduced how ribosomes are major determinants of bacterial cell growth rate. Similarly, as the growth rate increases, so does the volume of the cell (Schaechter, Maaloe, and Kjeldgaard 1958). Cell envelope associated pathways, for example those which produce peptidoglycan precursors for the cell wall, are responsible for surface growth. With the increase in cell volume must come an increase in surface area therefore requiring a concerted expression of all involved pathways with ribosome synthesis. Genes in conserved gene neighbourhoods and operons have been shown to co-regulate and provide reduced noise in their co-expression (Ray and Igoshin 2012). Therefore, the existence of a gene cluster containing rate-limiting genes for both cell volume and cell surface growth could synchronize these processes. To this end. we targeted 77 translation-associated genes, including ribosome subunits, translation factors and tRNA synthetases for genomic context analysis (Table 4.1). Only genes which were conserved over the significance threshold provided by GenCoDB were considered. 35 of these surveyed genes, 31 being ribosomal proteins, were part of the mega-ribosomal gene cluster. The mega-ribosomal gene cluster is the largest gene cluster found in nearly all bacteria and usually consists of the large majority of the genome's ribosomal protein genes as well as several other processes including ATP synthesis (Ohkubo et al. 1987). Of the other 17 ribosome proteins, all were found to be conserved with other genes. However, these were almost exclusively found with only other translation associated genes in much smaller clusters. Eight translation genes were found to have no significant conserved gene neighbourhoods. It was very rare that cell envelope genes were associated with translation genes. Only three of our surveyed genes were found in a genomic context with cell envelope genes which coincidentally all belong in the same cluster. This cluster contained three of the surveyed genes, *rpsB*, *tsf* and *frr*, which were found together with three cell envelope genes, namely, *uppS*, *cdsA* and *dxr* – all of which will be described in the next section in detail.

Genomic contexts of translation-associated genes

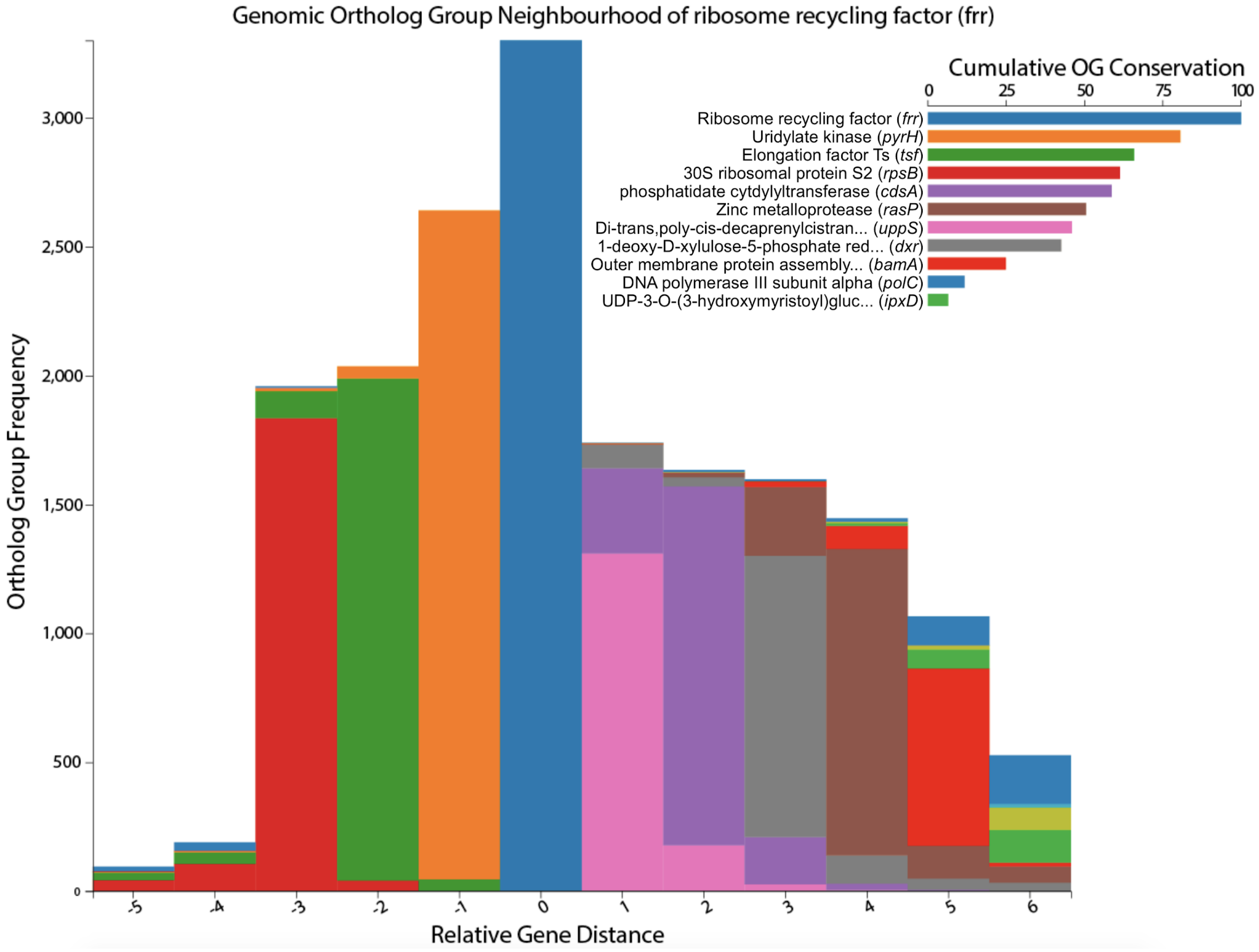
Gene name	OG ID	Ortholog group description	Conserved Neighbours	Ribosome-associated neighbours	Cell Envelope Neighbours
tsf	1405357at2	Elongation factor Ts	14	rpsB,frr	uppS, cdsA, dxr
rpsB	1623045at2	30S ribosomal protein S2	11	tsf,frr	uppS, cdsA, dxr
frr	134426at1385	ribosome recycling factor	11	rpsB,tsf	uppS, cdsA, dxr
yqeL	1990650at2	Ribosomal silencing factor RsfS	7	rplU, rpmA, YchF, RlmH	
pheS	469058at2	phenylalanine--tRNA ligase subunit alpha	6	pheS, rplT, rpmL, thrZ	
rplS	1698718at2	50S ribosomal protein L19	6	trmD, rpsP, RimM	
rplT	1932144at2	50S ribosomal protein L20	6	rpml, pheS, thrZ, infC	
thrZ	900765at2	threonine--tRNA ligase	4	rplT, rpmL, pheS	
rpmI	2046660at2	50S ribosomal protein L35	6	rplT, pheS, thrZ	
rpsP	1937072at2	30S ribosomal protein S16	6	RimM,rplL, trmD	
rpsR	1940575at2	30S ribosomal protein S18	4	rpsF, rplI	
rplI	1959318at2	Ribosomal protein L9	4	rpsR,rpsF	
rpsF	1776954at2	30S ribosomal protein S6	4	rpsR, rplI	
rpsO	1990141at2	30S ribosomal protein S15	7	rbfA, truB, infB, RimP	
infB	347113at2	translation initiation factor IF-2	8	rbfA, RimP, truB,rpsO	
rbfA	1971380at2	ribosome-binding factor A	9	RimP, infB, truB, rpsO	
prfA	928964at2	peptide chain release factor 1	1	rpmE	
rpmE	2014569at2	50S ribosomal protein L31	1	prfA	
rplU	1949059at2	50S ribosomal protein L21	4	rpmA	
rpmA	1904463at2	50S ribosomal protein L27	4	rplU, ychF	
rpmB	2062238at2	50S ribosomal protein L28	2	rpmGB	
rpmGB	2034291at2	50S ribosomal protein L33	3	rpmB	
gatB	1498741at2	glutamyl-tRNA amidotransferase	3	rpsU	
rpsU	2088764at2	30S ribosomal protein S21	3	gatB	
aspS	226836at2	aspartate--tRNA ligase	1	hisS	
defA	1649129at2	peptide deformylase	1	fmt	
glyS	213210at2	glycine--tRNA ligase subunit beta	1	glyQ	
hisS	277998at2	histidine--tRNA ligase	1	aspS	
rpmF	2092354at2	50S ribosomal protein L32	7		
metS	761140at2	methionine--tRNA ligase	1		
rpsT	2005443at2	30S ribosomal protein S20	1		
trpS	951354at2	tryptophan--tRNA ligase	0		
tyrS	402899at2	tyrosine--tRNA ligase	0		
valS	32262at2	isoleucine--tRNA ligase	0		
leuS	32262at2	isoleucine--tRNA ligase	0		
lysS	63621at2	lysine--tRNA ligase	0		
cspR	132510at1385	tRNA methyltransferase	0		
gltX	1409413at2	glutamate--tRNA ligase	0		
argS	1146366at2	arginine--tRNA ligase	0		

3 - Table 4.1 - Genomic context of translation associated genes

Survey of the genomic contexts for translation associated genes looking for the presence of co-localized cell envelope genes. Each column represents the number of ortholog groups (referred to as genes) found in the genomic neighbourhood. All bacterial genomes which contained the gene were considered and significance was determined by the default threshold provided by GenCoDB. In the case the gene fell into a well known cluster, the cluster name is provided instead. Genes found in the conserved ribosome cluster were omitted from the table.

4.2 The translation-cell envelope cluster

Based on its unique role as the only gene cluster associating genes involved in translation and cell envelope synthesis, we will refer to it as the translation-cell envelope (TCE) cluster. The cluster consists of eight core genes, the six already mentioned and two auxiliary genes being *pyrH* and *rasP*, which are involved in pyrimidine biosynthesis and cell division, respectively (Figure 4.1). We saw that the gene order of the cluster was conserved, matching the gene cluster behavior we observed in chapter 3, and is mostly seen in the order *rpsB*, *tsf*, *pyrH*, *frr*, *uppS*, *cdsA*, *dxr* and *rasP*. The gene with the strongest association with the cluster (based on the average co-localization of the other seven genes) was *frr* and the weakest was *dxr*. Because of the tight association between gene context and functional association it is important to understand the role of each of these genes/proteins in the cell to see how they potentially might interact. To this end we will briefly review each of the genes in this cluster.



30 - Figure 4.1 - The Translation Cell Envelope Cluster

A histogram representing the conserved neighbourhood around frr. The height of the bar represents the frequency that gene (ortholog group) appears in that position relative to frr.

Genes below the significance threshold are not displayed. The bars in the top right represent the sum each colour in the histogram as a percentage of the height of the center bar. As all genes displayed here are in the positive y-axis this means they are all in the forward orientation relative to frr.

Of the eight core conserved genes, ***rpsB*** is first gene of this cluster and is strongly conserved both in bacteria and in higher eukaryotes. *rpsB* encodes ribosomal protein S2 which is essential for the translation machinery in almost all cellular life. Although the functional role of S2 in translational activity is not yet completely understood, it is thought to play a key role in stabilizing the interaction of the ribosome with the Shine-Dalgarno sequence (Kaminishi et al. 2007; Yusupova et al. 2006) (Figure 4.2). In addition S2 may have moonlighting functions outside translation as it has been co-purified with RNA-polymerase and a global regulator *Hfq* (Sukhodolets and Garges 2003). Interestingly, S2 is specifically targeted by proteases in response to cell stress – potentially to both slow the cellular translation rate and to increase the free amino acid pool for new protein synthesis (Kuroda et al. 2001). Therefore, *rpsB* could act as a bottleneck for translation initiation, thereby dictating how fast the cell is growing under stress conditions.

The ***tsf*** gene encodes the translation elongation factor Ts (EF-Ts) protein. Its role in translation is to facilitate the dissociation of GDP from elongation factor-Tu (EF-Tu) so that EF-Tu may reform its active Ef-Tu-GTP complex which induces the binding of the codon specific aa-tRNA to the A-site of the mRNA-programmed ribosome (Gromadski, Wieden, and Rodnina 2002) (Figure 4.2). It has been shown *in vitro* that EF-Ts can inhibit RNA polymerase function (Biebricher and Druminski 1980). Therefore, it is possible that during stringent response when high concentrations of ppGpp and pppGpp block the interaction between EF-Tu and EF-Ts, the free EF-Ts would then be able to also restrict transcriptional activity.

Following *tsf* is ***pyrH*** which is an essential gene encoding uridylate kinase. Uridylate kinase plays its key role in pyrimidine biosynthesis by forming UDP through the ATP-dependent phosphorylation of UMP (Figure 4.2). UDP can then later be phosphorylated to UTP, an important substrate for RNA polymerase, a precursor for CTP synthesis and a cofactor in sugar metabolism (Figure 4.2). The production of UDP must be tightly regulated as it can be readily transformed into dUDP and then dUTP, which could be incorporated into DNA instead of dTTP. This is why it is thought that *pyrH* is localized to the periphery of the cell (Noria and Danchin 2002). Uridylate kinases are feedback-inhibited by UTP and activated by GTP (Gagyi et al. 2003), suggesting a homeostasis mechanism to ensure balanced purine and pyrimidine metabolite pools (Figure 4.2). The mechanism in which GTP activates uridylate kinases is different from bacteria to bacteria and does not appear to extend to archaea. Downstream, UDP-Glc has been shown to be a negative regulator of FtsZ assembly (Weart et al. 2007) and that low levels of UTP can lead to the modulation of attenuator sequences (Bonner et al. 2001).The former means that control of UDP and central carbon metabolism link growth rate to cell division to ensure they occur together in a growth rate dependant manner. The latter allows the cell to increase pyrimidine biosynthesis in events of low UTP levels by promoting the formation of an antiterminator before the pyrimidine biosynthesis gene cluster and potentially other genes (Bonner et al. 2001).

frr is the next gene in the gene cluster, encoding the ribosome recycling factor protein. It is responsible for the dissociation of ribosomes from mRNA after the termination of translation so that all factors (ribosome, mRNA and tRNA) may be recycled for the next round of translation (Janosi, Shimizu, and Kaji 1994) (Figure 4.2). This mechanism is in conjunction with an association to elongation factor G and the hydrolysis of GTP (Figure 4.2). Overexpression of the *frr* gene has been shown in *Streptomyces diastatochromogenes* to lead to increases in both cell growth and protein production (Ma et al. 2014).

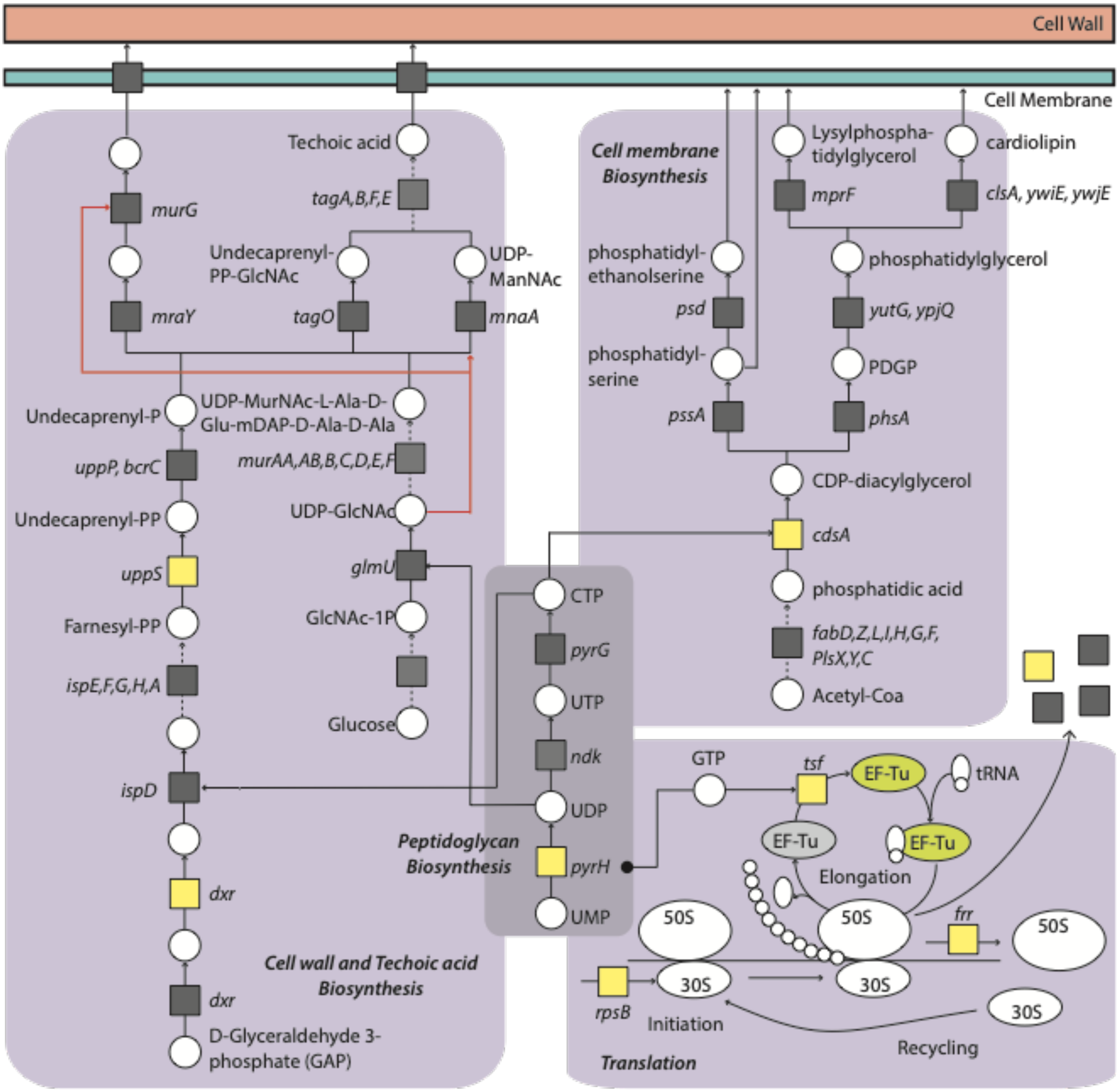
uppS encodes the cis-prenyltransferase undecaprenyl-diphosphate synthase (UppS) which is an important enzyme in the synthesis of the bacterial cell wall. The protein catalyses the elongation of C15-PP with eight isoprene units, creating di-trans, octo-cis-undecaprenyl-diphosphate (UPP, C55-PP) which can enter the lipid II cycle by being dephosphorylated to undecaprenyl phosphate (Figure 4.2). This is then enzymatically converted to lipid I and lipid II by the subsequent addition of amino acid-bound and activated UDP-sugars, forming a peptidoglycan precursor. Lipid II acts as a carrier to flip the peptidoglycan precursor across the membrane where they can form part of the growing peptidoglycan chain. Therefore, *uppS* can directly control the number of carrier molecules in the Lipid II cycle. Furthermore, the undecaprenyl phosphate is used in *B. subtilis* by TacO in the biosynthesis of wall teichoic acids. Wall teichoic acids are a major component of the cell wall in gram-positive bacteria (comprises 60%) and if they are removed result in defects in cell division and cell morphology (Brown, Santa Maria, and Walker 2013). Therefore, ensuring critical numbers of wall teichoic acids is essential for normal cell growth.

cdsA encodes a phosphatidate cytidylyltransferase which forms the key lipid intermediate cytosine diphosphate-diacylglycerol from phosphatidic acid and cytosine triphosphate. This is the last step in lipid biosynthesis before the diversification of the different polar head groups and is the main source of this key membrane precursor (Figure 4.2). *E. coli* mutants with low levels of CdsA activity have been shown to have no growth rate defects, however, this was accompanied by modifications in the lipid composition of their membranes (Ganong, Leonard, and Raetz 1980). Aside from forming the envelope, the phospholipid membrane has been shown to be loosely coupled to membrane protein homeostasis, as protein synthesis and growth has been shown to be halted until the phospholipid to protein ratio is restored (McIntyre et al. 1977).

dxr encodes 1-deoxy-D-xylulose 5-phosphate (DXP) reductoisomerase. DXP reductoisomerase is the first enzyme devoted to the MEP pathway, one of the two pathways involved in isoprenoid synthesis (Figure 4.2). Isoprenoids are a diverse group of molecules found in all organisms with a range of functions and are being considered an alternative biofuel to ethanol. Whilst the function of isoprenes in bacteria is not fully known, many of the genes in this pathway are essential for normal growth and many precursors within the pathway are toxic (Sivy, Shirk, and Fall 2002). In bacteria they can serve a role as a cell wall biosynthesis intermediate. Overexpression of *dxr* was shown to not increase the production of isoprenes (Xue and Ahring 2011). If not diverted towards isoprene production, the downstream product of *dxr* eventually acts as a precursor for *uppS*, discussed earlier. In fact both *uppS* and *dxr* were shown to be negatively correlated with isoprene production and may instead indicate a shift away from isoprene production towards larger terpenoids (Hess et al. 2013).

The last gene in this gene cluster is ***rasP*** which encodes an intramembrane protease. As a protease, it degrades pre-cleaved targets, several of which are signal peptides. In specific relevance to cell growth in *B. subtilis* it is involved in the degradation of RsiW, an anti-sigma factor to σ_w (Parrell et al. 2017). σ_w is a regulator that acts in response to cell envelope stress (Zweers et al. 2012). In overexpression strains of *rasP*, it was shown to boost the production of several membrane proteins

(Neef et al. 2017) presumably by clearing the membrane of unneeded or mislocalized proteins that would perturb the cell envelope. Another important function is that it cleaves the essential cell division protein FtsL. FtsL has been shown to be rate limiting for cell division (Bramkamp et al. 2006) so it is speculated that a high expression of *rasP* promotes cell elongation over division (with cell wall stress being stimulated by the rapid volume growth). Given this finding, and its context with several ribosome-associated and cell envelope genes, *rasP* could be an important regulator of cell size and division in relation with the growth rate



31 - Figure 4.2 - Cluster genes hold bottleneck positions

The cell envelope associated genes hold bottle-neck positions in their respective biochemical pathways. These steps are often before committed steps. RpsB is a required protein for translation initiation as a part of the ribosome complex. Tsf is a bottleneck for translation elongation reactivating Ef-Tu (inactive in grey, active in green). PyrH is the sole producer of UDP in *de novo* pyrimidine biosynthesis. frr controls termination of translation and therefore the available active ribosome pool. Together RpsB, Tsf, and Frr contribute to translation producing more proteins, including their self-replication. UppS produces the carrier molecule for the lipid II cycle and is removed from the cluster in species where it has been duplicated and therefore made redundant. CdsA produces the last intermediate before the diversification of the polar head groups. Dxr is an essential gene in teichoic acid biosynthesis and produces an intermediate required for cell wall synthesis. Squares represent enzymes and are coloured yellow when they belong to the TCE cluster. Circles are metabolites.

Selective forces of the TCE cluster

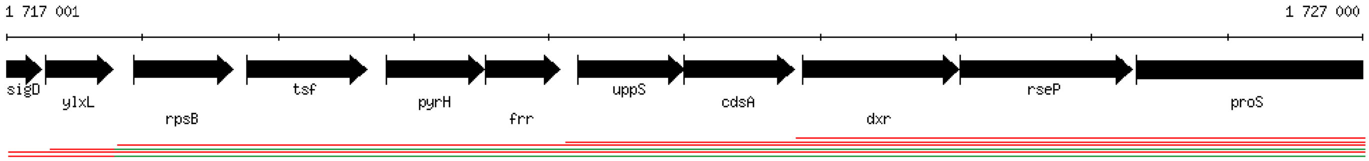
To understand what evolutionary benefits co-localizing these eight genes together would provide to a species, we analysed it under the three frameworks of conservation that we developed in Chapter 3: essentiality, protein-protein interactions within the cluster, and operon organized transcription. Essentiality results in genes less likely to be deleted or displaced (disrupting expression) resulting in the grouping of essential genes as non-essential genes in between get removed. Excluding *rasP*, all the core conserved genes have been demonstrated as essential in many organisms (Luo et al. 2014) and all are highly conserved across the Bacteria kingdom. The high number of essential genes in this cluster means that it is extremely unlikely that the broad dispersal pattern of the cluster was due to horizontal gene transfer and therefore was most likely transmitted vertically. It should be noticed that several essential ribosome associated genes are not found in conserved clusters (Table 4.1) and unlike other clusters where essentiality of the genes is enriched, the cluster comprises many different processes. Therefore, we believe it is unlikely that the only selective force keeping this cluster together is essentiality. Proteins involved in within-cluster protein-protein interactions have been frequently observed to be encoded in gene clusters. It is thought this is advantageous as coupled transcription and translation results in increased local concentrations of the co-localized proteins, increasing the probability of finding interaction partners. Despite this, there is very little evidence in the literature detailing interactions between the eight proteins encoded in the cluster, aside from Frr associating with the assembled ribosome (and thereby indirectly with RpsB) at the end of translation. As many of the genes are from different pathways there is no intuitive explanation as to why they would need to interact but that does not preclude moonlighting interactions.

In chapter 3 we found that operon level transcription, whilst likely not a driving mechanism for cluster formation, was highly abundant in gene clusters. Unlike eukaryotes, bacteria have a mechanism in which to ensure equal transcript abundances. Operons were first discovered by (Jacob and Monod 1961) and involve the transcription of multiple genes from one promoter resulting in a transcript encoding multiple genes (polycistronic transcripts). Evidence that many of the genes in this cluster are transcribed as polycistronic transcripts include that in practically all genomes in which the cluster is maintained, the genes are co-orientated (Figure 4.1). From our analysis in Chapter 3, we found that this cluster was enriched in operons in *B. subtilis*. However, this calculation measures how many genes are found in an operon with at least one other gene from the cluster and does not take into account the number of genes that are grouped together. Therefore, from this observation alone, we cannot know if the entire cluster is found in an operon together and how strongly the co-regulation between genes is. Unlike many operons, the genes in this gene cluster are involved in very different biological processes and yet they are unified in that they each hold either bottleneck or rate limiting steps in these respective processes (Figure 4.2). Even the three genes involved in translation belong to separate different rate limiting steps in translation: initiation, elongation and dissociation. In the introduction we have already discussed the requirement of synchronization between both surface and volume expansion, or interpreted as cell envelope synthesis and translation/ribosome content. Therefore, it would be highly beneficial for a cell to tightly regulate these processes through a single operon to ensure synchrony.

4.3 Regulation of the cluster

Literature suggests polycistronic transcripts across the gene cluster

Microarray data from *B. subtilis* compiled by Subtiwiki (Zhu and Stülke 2018) have suggested that there is operon level regulation occurring over this cluster (Figure 4.3), as indicated by one long transcript spanning from *rpsB* until downstream of *rseP (rasP).* We should however view these data with caution as microarray-based studies are limited in their ability to define operons. For example, due to the low base-pair resolution of microarray data it is possible that neighbouring transcriptional units would be detected as operons even if separated by strong terminator and promoter elements. If they are indeed transcribed as an operon, from these data we cannot quantitatively distinguish between polycistronic and monocistronic transcripts. Finally, this microarray observation also conflicts with the operon databases we used in chapter 2, which currently do not classify all the genes as forming a single transcript (Mao et al. 2009). Therefore, it became clear to us that we needed to further investigate the exact transcriptional organization and regulation of this gene cluster.



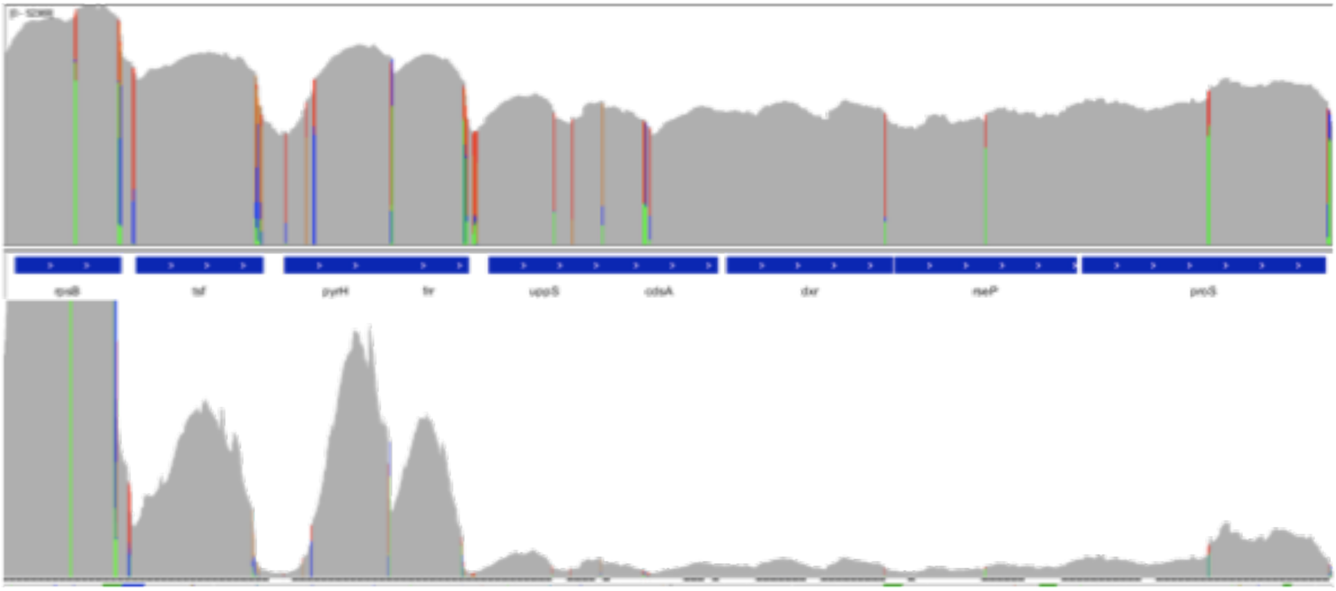
32 - Figure 4.3 - Microarray data showing operon level organisation in *B.subtilis*

Transcriptional units for the TCE cluster predicted with microarray data. The red lines each represent independent transcriptional units. Figure was acquired from Subtiwiki (Zhu and Stülke 2018) and altered. The existence of one long polycistronic transcript containing all eight genes of the cluster is suggested.

Absolute transcript levels varying between the 8 genes

To observe how these genes are regulated in these species and in various conditions, we downloaded several RNAseq datasets for model organisms from the four main phyla in our dataset, *B. subtilis, E. coli, M. tuberculosis and B. fragilis* (Firmicutes, Proteobacteria, Actinobacteria and Bacteroidetes respectively) (for conditions see Table 8.9). The TCE cluster is present all of these species except for *B. fragillis* where the genes have been dispersed across the genome. The definition of an operon requires that all member genes would be transcribed at the same abundance under the assumption of no internal transcription start sites (TSS) or terminators (aside from abortive transcription resulting in the progressive reduction of expression along the cluster length). Even in cases where the assumption of no terminators is invalid, transcript levels should remain correlated across different conditions. We found that transcript levels varied significantly between the genes suggesting the presence of regulatory elements within the cluster (Figure 4.4). The expression of the cluster was highest for genes at the 5' end of the gene cluster and decreasing for genes at the 3' end highlighting the role terminators play in the gene regulation of this cluster.

Strong upshifts of transcripts within the cluster indicate the presence of internal transcript start sites.

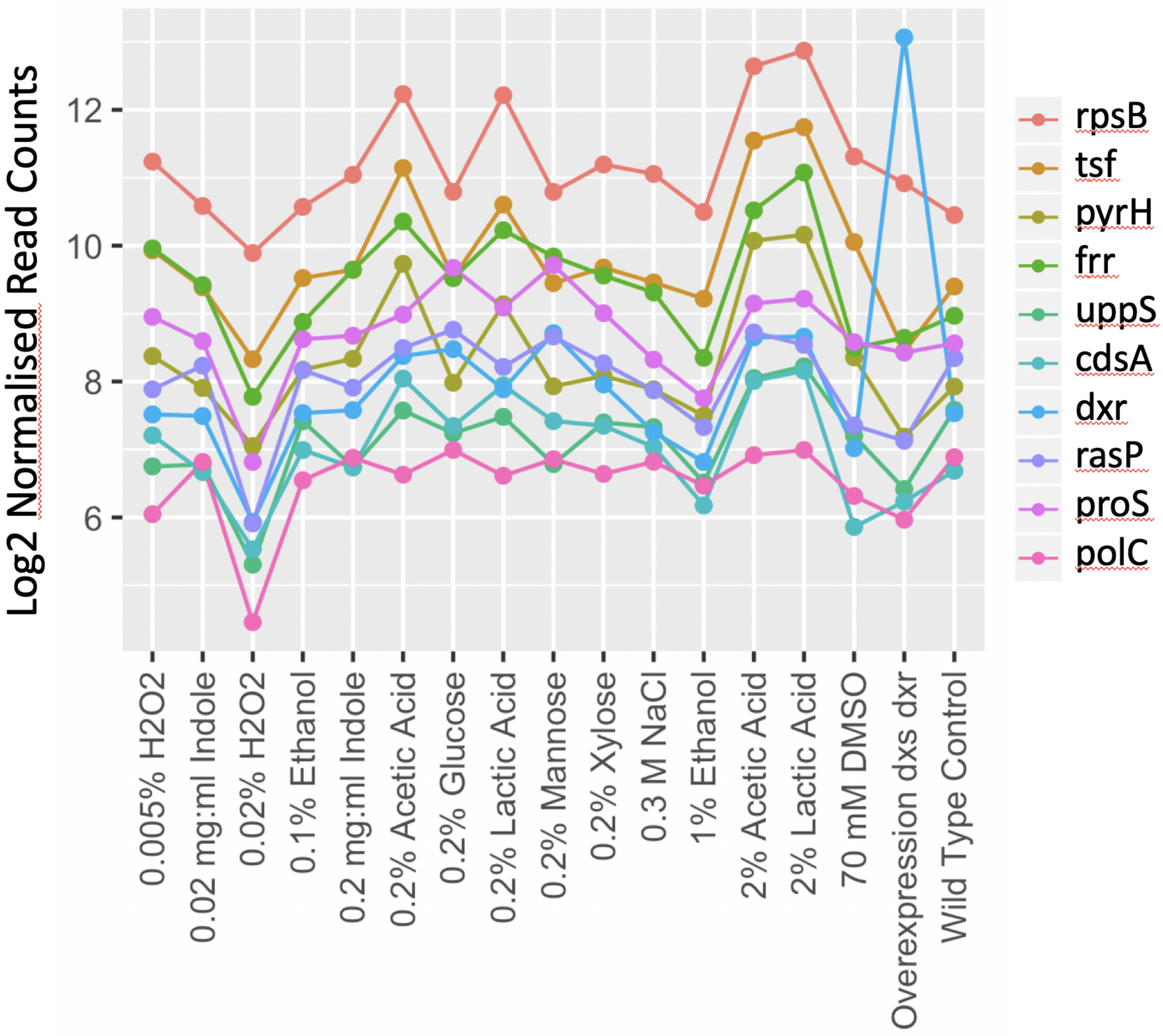


33 - Figure 4.4 - The transcriptional profile of the TCE gene cluster in *B. subtilis*

The transcriptional profile over the TCE cluster of *B. subtilis* grown in LB media during exponential phase. The height of the grey area represents the relative number of transcripts mapping to the genomic region containing the TCE cluster. Genes on the chromosome are represented in blue. The upper and lower profiles are displaying the same data. The upper profile has been log transformed and the lower profile has had the y-axis truncated to display the less expressed genes in a linear scale. Viewing the transcriptional profile on a log scale hides the vast changes in transcriptional activity that occur over the cluster. Coloured lines represent single nucleotide polymorphisms between the sample and the reference genome. Large upshifts represent transcription start sites and conversely downshifts suggest locations of transcriptional terminators.

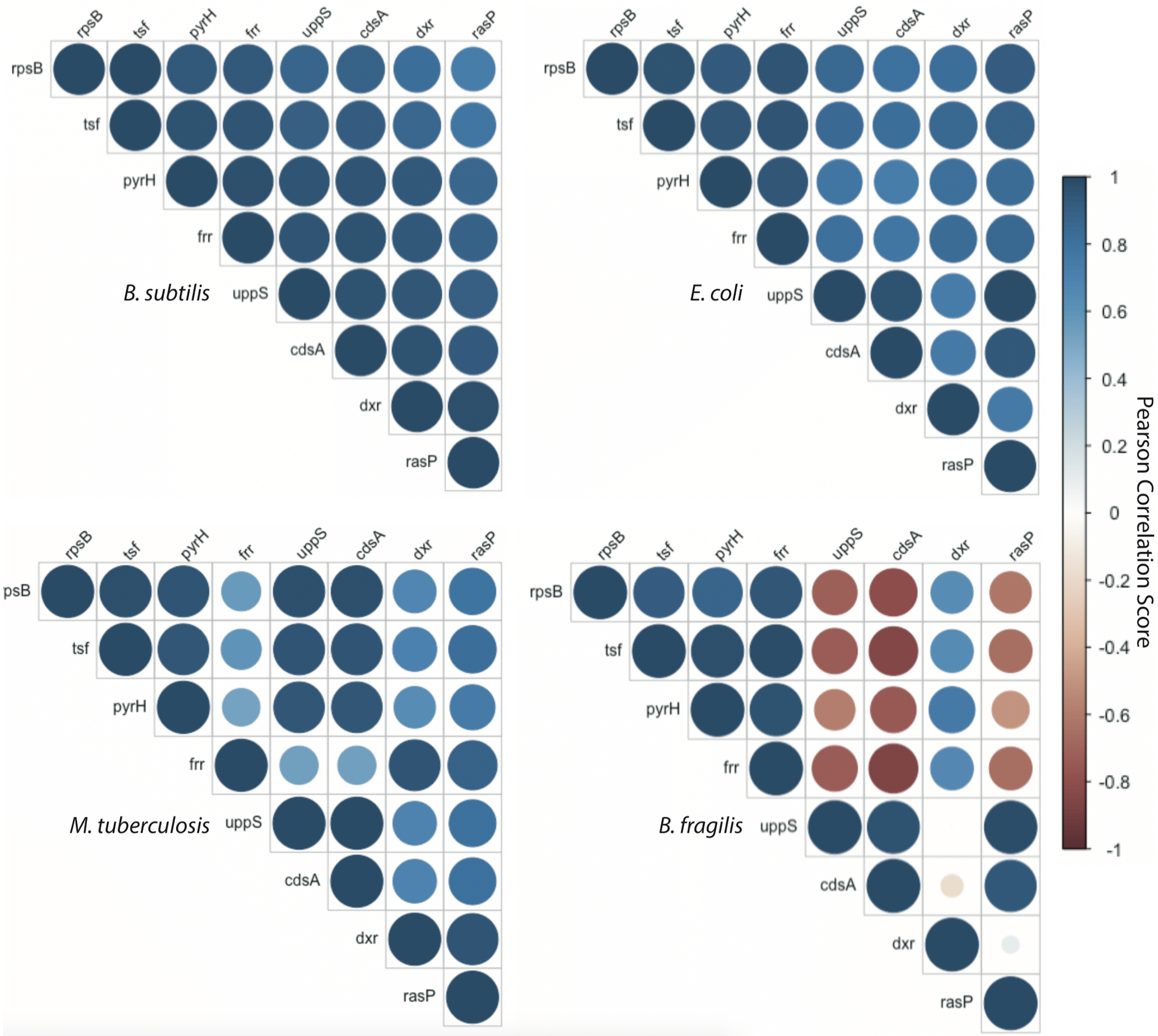
Gene expression of the cluster correlates across conditions and species

The apparent presence of internal transcript start sites challenges the idea that these genes are transcriptionally synchronized as they could all be independently regulated. We therefore took the available RNAseq datasets and measured the correlation of gene expression of the TCE cluster genes in several different condition (Table 8.9). Despite the seemingly high presence of promoters and array of different conditions, the expression of the eight genes in *B. subtilis* correlated strongly between the conditions, as indicated by visual inspection of the transcriptional profile (Figure 4.5*)*. To quantify this, we calculated Pearson correlation scores, where positively correlated genes will have scores close to 1, and genes which do not correlate at all will have scores close to 0. Negatively correlated genes will have a score close to -1. Of the three species where the cluster was conserved (*B. subtilis*, *M. tuberculosis*, and *E. coli*) we saw high correlation scores (>0.9) suggesting these promoters may be regulated by the same mechanisms (Figure 4.6). Conversely, we saw a loss of positive correlation between the genes on the *B. fragillis* genome where the cluster is split (Figure 4.6). However, as we measured in chapter 3, average correlation scores for gene expression in neighbouring genes is very high, this cluster would not be classed as significantly correlated by this stringent threshold (Figure 3.15). When compared to the expression correlation of random genes we do see the cluster is higher than the stringency threshold.



34 - Figure 4.5 - Tightly correlated expression of the cluster genes in diverse conditions

A selection of conditions from the RNAseq dataset retrieval of *B. subtilis* showing number of reads for each of the eight cluster genes and two additional genes found conserved in Firmicutes. Expression of the cluster genes remains highly consistent in all the observed conditions regardless of perturbation or carbon source (growth rate). The exception here is *dxr* which had higher levels of *dxr* expression in a *dxr* overexpression strain (This condition was not included in future expression correlation analysis).

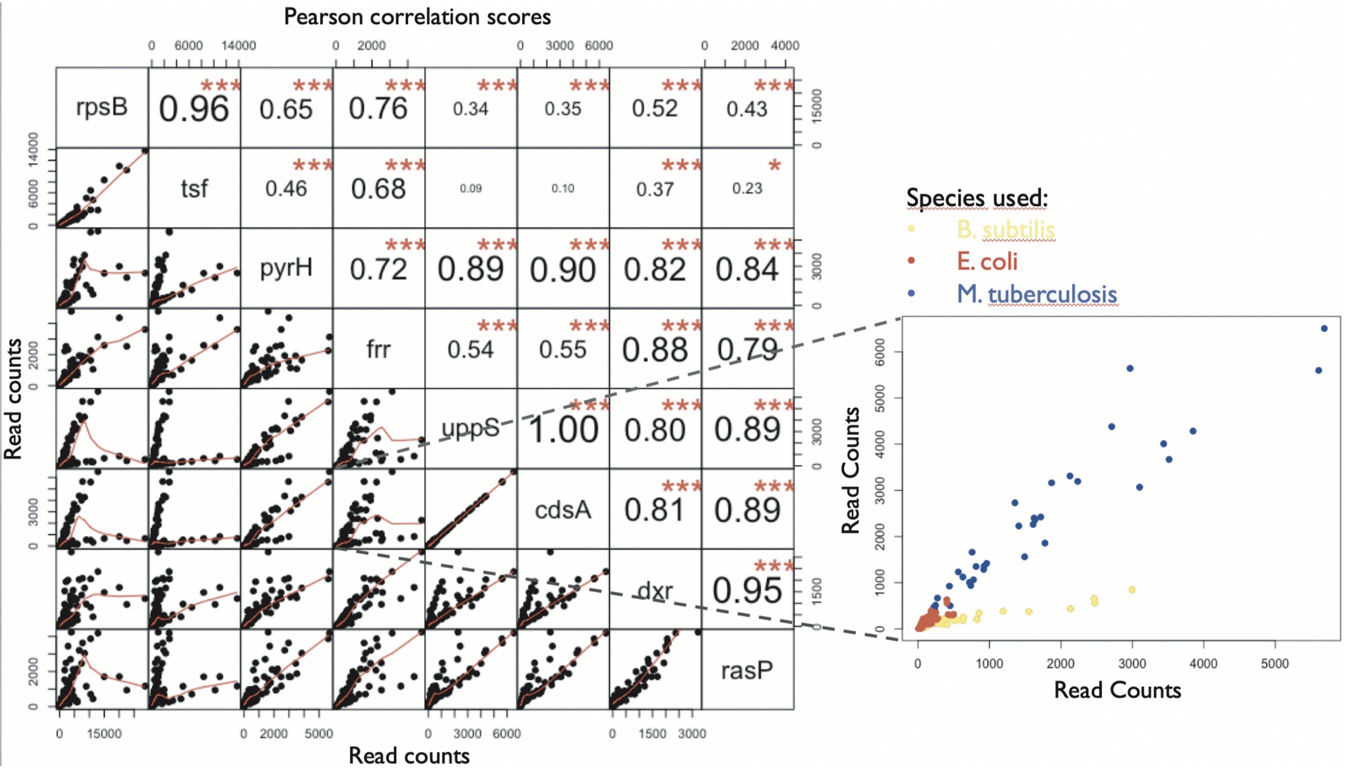


35 - Figure 4.6 - Pairwise correlation of TCE expression in four different Bacteria species

Pearson correlation tests of the expression between the pairwise comibinations of the eight cluster genes across different conditions in different species. The exact retrieved conditions varied between the species (n = 62, 72, 58, 30 for *B. subtilis*, *E. coli*, M. tuberculosis, *B. Fragilis* respectively) based on the available datasets. The colour and size of the circle represents the Pearson correlation score taken from all conditions of that sample.

Ratio between cluster genes is not conserved between species and highlights the requirement of fine tuning within the cluster

To see if the presence of internal transcription start sites allows for different species to fine tune the expression of the cluster genes, we pooled together the datasets from the three cluster-containing species and reperformed the correlation analysis. Interestingly, whilst within species correlation between the gene cluster's transcript expression is conserved, this is lost in many gene pair comparisons between species (Figure 4.7). We found that the loss was due to different ratios of expression between the different cluster genes of the different species. This would represent the need for different species to have different levels of the proteins to accommodate their growth needs. Altogether this can explain the presence of the internal transcription start sites as they may act as fine tuners, setting the absolute requirement of the gene product for the species' physiological needs. Interestingly however, *uppS* and *cdsA* expression levels were tightly correlated across all three species, which is unexpected, given the increased peptidoglycan needed in gram-positive species, and the increased phospholipids needed in gram-negative species.



36 - Figure 4.7 - Pooled pairwise read expression levels of the cluster

RNAseq datasets between the three cluster containing species were pooled to observe inter-species correlation of the TCE cluster genes. The numbers in the upper triangle reflect the Pearson correlation score. The number of stars represent the p-value. * < 0.05, ** < 0.01, * < 0.001. The lower triangle shows the expression of each gene plotted against each other with each dot presenting a different RNAseq sample. The inset shows a zoomed in view at the f*rr-cdsA* comparison. The points have been coloured based on the species the sample was derived from (blue = *M. tuberculosis*, red = *E. coli,* yellow = *B. subtilis*)

Confirmation of several transcription start sites within the gene cluster

The differing stoichiometries in the TCE cluster gene levels between species suggest there must be regulatory elements within the cluster to module expression of the individual genes. We wanted to understand exactly where and how many regulatory elements were found in the gene cluster. The upshift in transcript abundances were most apparent before *tsf*, *pyrH* and *frr* and the reduction of transcript abundance happened directly after the first four genes of the cluster. To find the transcription start sites we performed 5' RACE experiments on extracted RNA from *B. subtilis. B. subtilis* was grown to exponential phase in LB media for RNA extraction. After reverse transcription and polyadenylation of the cluster genes (see Materials and Methods) we were able to confirm the presence of eight different start sites. The sites were mainly clustered in the first half of the gene cluster, namely: one before *rpsB*, one before *tsf*, three before *pyrH*, one before *frr*. Surprisingly, we were unable to find a transcription start site for *uppS* which also did not have a clear spike before in the transcriptional profile of the cluster (reference to Figure 4.4).

TSS position (bp)	Subsequent Gene
-52	*rpsB*
782	*tsf*
1180	*pyrH*
1584	*pyrH*
1843	*pyrH*
2097	*frr*
3568	*cdsA*
5900	*rasP*

4 - Table 4.2 - 5' RACE identified transcription start sites

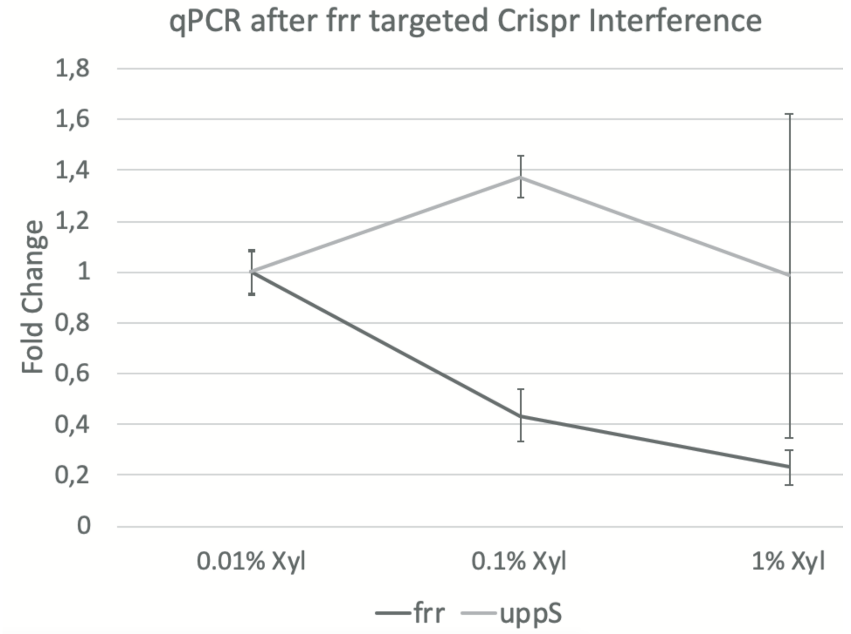
Predicted transcription start sites (TSS) acquired from 5' RACE experiments on extracted RNA from *B.subtilis* grown in exponential phase in LB medium. The TSS position was measured from the first "A" from the start codon of *rpsB*. The subsequent gene represents the first gene which has a start codon after the identified TSS.

Whole cluster transcripts unlikely due to the presence of strong terminators

During transcription, terminators are not complete roadblocks for RNA polymerase and have been shown to have varying levels of efficiency (Mitra et al. 2009). In fact many terminators can be regulated to modulate this efficiency, which can in turn stop or increase the level of polycistronic transcripts (Schmidt and Chamberlin 1987). Therefore, the possibility remained that there was a basal level of transcription that started at the *rpsB* promoter and continued transcription until the end of the cluster. If this existed, one could posit that this fraction of long transcript is the basal level required for the synchronisation of the genes at a specific growth rate. From re-analysis of the data published by (Mondal et al. 2016) we noted three terminators, which reduced transcriptional read through by 75%, 81% and 60% after *rpsB*, *tsf*, and *frr,* respectively (Figure 4.4). It was clear that the chance for concurrent transcription from one end of the gene cluster to the other is extremely low. Given the abundance of internal transcription sites and terminators within the cluster we deemed it unlikely that operon level transcription was relevant in the expression of these all these genes. However, the absence of a detectable transcription start site before *uppS* with the 5'RACE method suggests *frr* could be transcriptionally co-transcribed with its downstream genes, such as *uppS* and *cdsA* with partial transcription abortion through the relatively weak terminator after *frr*. We cannot however preclude the possibility our method was unable to detect a TSS in this region.

No significant readthrough occurs between translation and cell envelope parts of the cluster

Due to the absence of a detectable TSS and a relatively weak terminator element between *frr* and *uppS* we sought to measure how many *uppS*-containing transcripts derived from upstream expression. As PCR over this area could only confirm if the long transcripts existed and not quantitatively measure if upstream transcription from *frr* was the major transcript of *uppS*, we implemented a CRISPR interference (CRISPRi) system (Peters et al. 2016). CRISPRi, through a deactivated Cas9 (dCas9) which is guided to the DNA for a guide RNA and sterically blocks oncoming RNA polymerases, causing the polymerase to dissociate from the DNA, which results in an incomplete, non-functional transcript. The dCas9 does not bind to the target site permanently, thereby allowing RNApol progress between binding events. The level of repression can be controlled by the concentration of guided dCas9 proteins resulting in the occupation time of the targeted location. We used CRISPRi to knockdown RNA transcription of *frr* (Peters et al. 2016). To ensure high efficiency knock-down and to reduce the possibility of interacting with an internal promoter within *frr*, which could drive *uppS* expression, the guide RNA was targeted at the 5' end of *frr*. We expected that if *uppS* and *frr* were found on the same transcript, knocking down *frr* transcripts should result in a proportionally similar knockdown of *uppS* transcripts. This however was not the case, as we saw that knocking down *frr* transcripts resulted in no correlated change in *uppS* transcript abundance (Figure 4.8). This shows that the expression of *uppS* and subsequently the downstream cell envelope genes are independent of the expression being driven from the upstream promoters.



37 - Figure 4.8 - qPCR after frr targeted Crispr Interference reveals no readthrough from frr transcripts

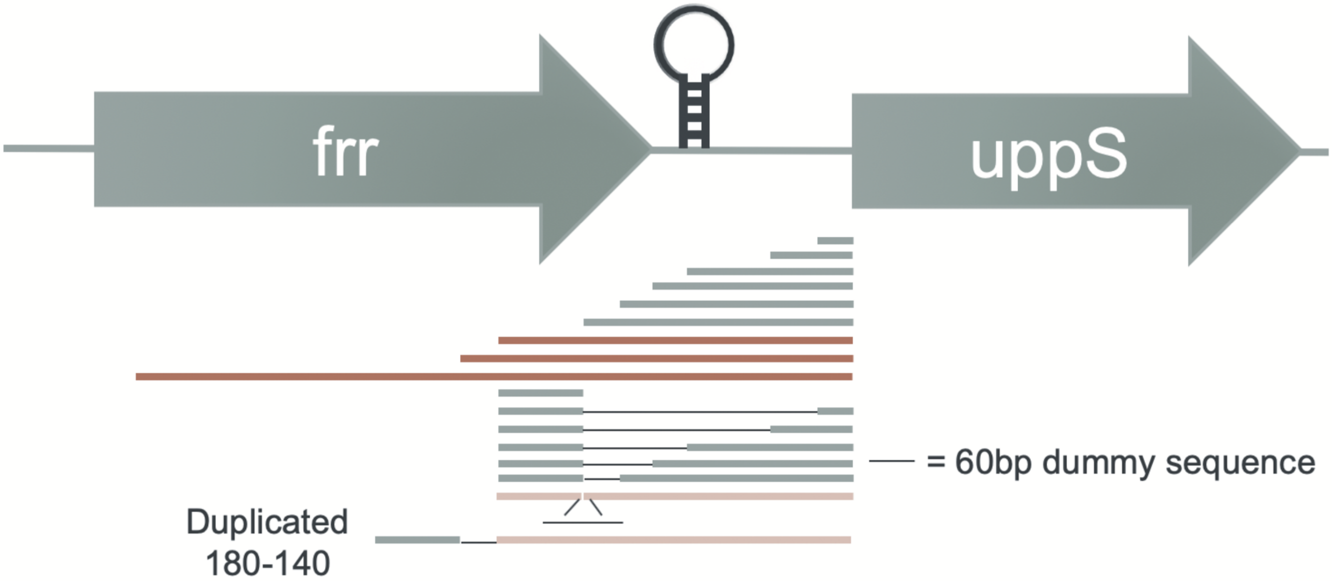
RNA was extracted from *Bacillus subtilis* mutant expressing dCas9 with a sgRNA targeting *frr* and grown in LB with varying concentrations of xylose and subjected to qPCR. Higher concentrations of xylose represent an increased knockdown of *frr*. To see how many *uppS* transcripts contained the *frr* coding sequence both *frr* and *uppS* transcript levels were measured with qPCR. Error bars represent the standard error between the replicates. Fold changes are relative to the transcript abundances measured in the 0.01% xylose condition. qPCR levels were normalized to the constitutively expressed genes *recA* and *gyrB.*

4.4 Characterization of the uppS promoter

uppS promoter detected within the frr gene

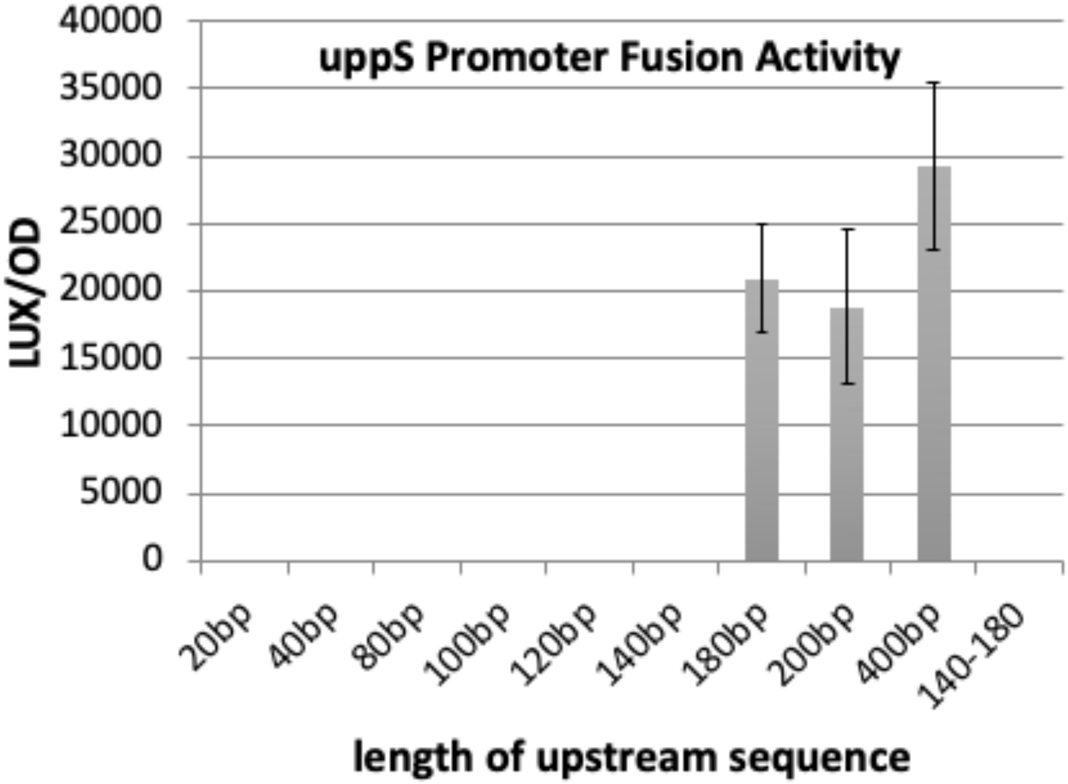
Given the presence of a terminator directly before *uppS* and no significant read-through occurring through this terminator from *frr*, the evidence strongly pointed towards an independent *uppS* promoter, despite not detecting a TSS with 5'RACE. Therefore, we created promoter fusions of the upstream area of *uppS* fused to the luciferase operon. We tried several 5' truncations of the genomic sequence before *uppS* and found that we needed 180bp before we observed reporter activity (Figure 4.9). Taking a larger fragment of 400bp, containing the majority of the *frr* gene, does not increase expression significantly, suggesting the presence of only one active promoter element (Figure 4.10). The next tested smaller fragment containing the first 140bp of upstream sequence provided no activity, suggesting the 40bp difference between these fragments were essential for activity and most likely contained the promoter (Figure 4.10). However, the 40bp alone did not show promoter activity. Combining it with increasingly sized parts of the UTR of *uppS* did not restore activity until the full 180bp was present. Due to the cloning method used, there was a 40bp between the 40bp and the 140bp fragment (Figure 4.9) and we believe this resulted in a reduction in activity (Figure 4.9). This could suggest that the promoter actually lies in the region 160-120 bp upstream of *uppS* and was incomplete in the 140bp fragment and therefore also truncated in the 40bp part explaining why there was no detectable activity from either of the promoter fusion constructs alone. Additionally, duplicating this 40bp region before the full 180bp region resulted in a reduction in activity of this putative promoter (Figure 4.9). This conflicts with the previous explanation and could suggest there is some repressive binding activity occurring from this sequence or that it is titrating away transcriptional activators.

One of the largest UTRs in the cluster in *B. subtilis* (130bp), exceeded only by the UTR between *tsf* and *pyrH* (146bp), lies between *uppS* and *frr*. As activity only occurs in fragments larger than this UTR this suggests the promoter may fall within the *frr* gene itself. It is striking that within the UTR, and therefore downstream of this putative promoter, there is a terminator with an efficiency of 60%, as mentioned previously. This means that there could be wasteful transcription initiation which is terminated shortly after beginning. However, these shorter transcripts may not form correct hairpin structures and allow transcription through this terminator. This may explain why the terminator was measured with only 60% efficiency despite us seeing little to no readthrough with the CRISPRi experiments (4.8). The presence of the terminator following the TSS start site may explain why we were unable to detect a TSS with the 5' RACE method as RNA secondary structures can reduce the efficiency of reverse transcription or later polyadenylation, thereby resulting in no primer binding site for subsequent PCR amplification and detection.



38 - Figure 4.9 - Schematic of promoter fusion constructs

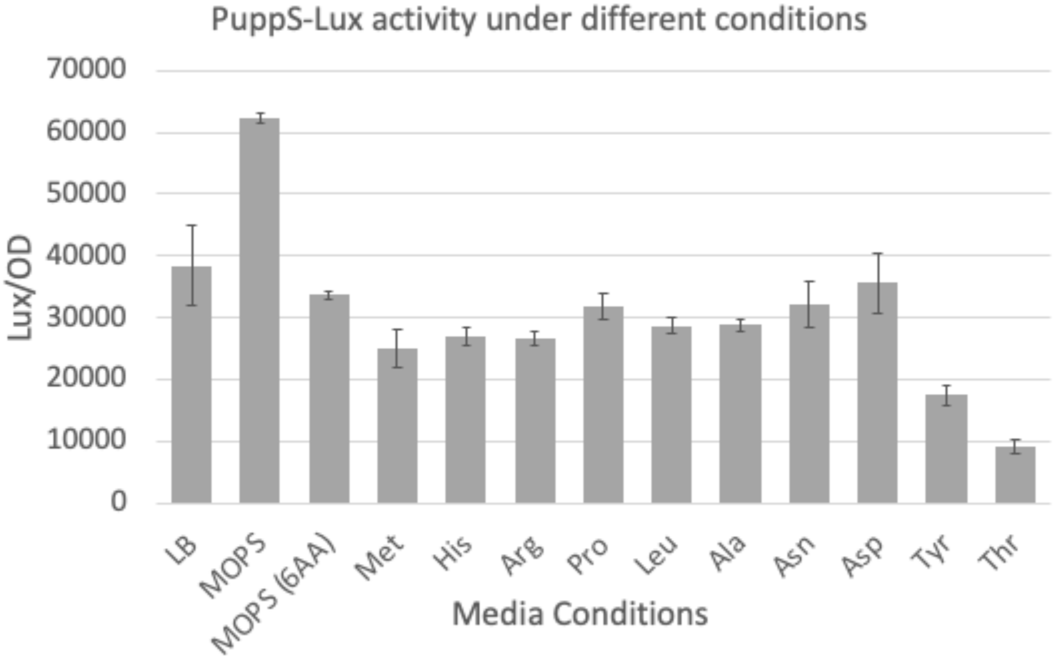
Construction of different promoter fusions to identify the active transcription region driving uppS expression. Each bar represents the length of the different tested promoter fusion constructs under their respective derived sequence. Grey lines represent inactive constructs, red lines represent active constructs and the faded-red constructs represent active constructs, however with weaker activity. Thin lines represent a 60bp dummy sequence that is introduced due to the cloning strategy and contains synthetic transcriptionally inactive DNA sequence used to connect the different 5' UTR regions.



39 - Figure 4.10 - Reporter activity of uppS 5' UTR truncations
Different truncations of sequence upstream of *uppS* were placed before the LUX reporter and integrated into the *Bacillus subtilis* genome. The length is measured from the 10bp upstream of the uppS start codon. Measurement was taken during exponential phase in LB media. Error bars represent the standard deviation between the replicates (n=3). 140-180 represents that only the sequence difference between the 180bp and 140bp fragments was used in the promoter fusion construct.

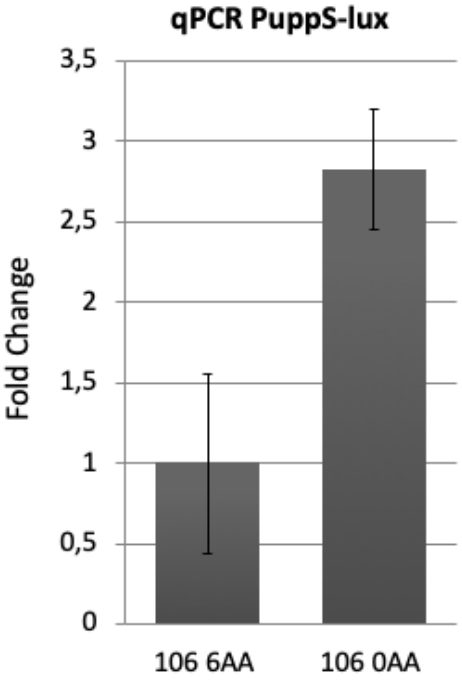
P*uppS* is transcriptionally upregulated in low amino acid conditions

Now that we identified the promoter which independently controls *uppS* we wanted to understand how this promoter was regulated. Given its tight genomic association with ribosomal genes and the need for increased lipid II carriers at higher growth rates we would expect the promoter to be upregulated under faster growth conditions. We measured the activity of the promoter in fast and slow growth conditions, respectively LB and MOPS media which in our plate reader promote doubling times of ~23 and ~78 minutes. Surprisingly, we found that opposite to what we expected, the activity of P*uppS* doubled in the slow growth conditions (Figure 4.11). To confirm this relationship, we tested an intermediate growth rate using MOPS media supplemented with amino acids (doubling time ~60 minutes), however, unexpectedly we saw the same level of reporter activity as with the much faster LB media (Figure 4.11). Supplementation of any of the individual amino acids used in the 6 amino acid mixture (methionine, histidine, arginine, proline, threonine) or any other amino acids we tested resulted in similar activity levels seen in the LB media (tryptophan the other amino acid is always present in the media). We did see reduced activity from supplementation of tyrosine and threonine but these both had detrimental effects on the growth rate of the cell. We cannot confirm the stimulus is slow growth in this case as between the MOPS media and LB media, there are many other nutritional differences. Given that we saw similar reductions upon the supplementation of all tested amino acids, a likely reason would be the nitrogen scavenging from the amino acids directly. Promoter fusion reporter assays measure expression activity of a nucleotide sequence, however, this could be derived from both transcriptional regulation (the promoter) and post transcriptional regulation (such as proteins binding to the 5'UTR). Therefore, we performed qPCR analysis of our exponentially grown 180bp-P*uppS* strain in both MOPS media with and without the amino supplementation to measure if the transcript levels of the first gene in the luciferase operon increased. We saw a similar two-fold increase in the MOPS media condition, suggesting that regulation of the *uppS* promoter occurs by a transcriptional mechanism (Figure 4.12). As *uppS* is a key gene in the production of peptidoglycan and wall teichoic acids, we hypothesized that this additional promoter allows the cell to respond to perturb ants of these pathways by increasing production of cell wall intermediaries through the expression of *uppS*. Preliminary experiments, testing sub-lethal concentrations of four different cell wall targeting antibiotics which trigger the σ^M response, bacitracin, nisin, ramoplanin and tunicamycin, showed no effect towards the activity of the promoter.



40 - Figure 4.11 - P$_{\text{uppS}}$ reporter fusion activity under different conditions

The 180bp P*uppS* fragment activity was measured under different media conditions. The measurement was taken during exponential phase. Glucose and tryptophan were added to all MOPS media conditions. MOPS media (6AA) include methionine, histidine, arginine, proline, threonine in the media. The other amino acid conditions are each individually added to basic MOPS media. Error bars represent the standard deviation between the replicates (n=3).



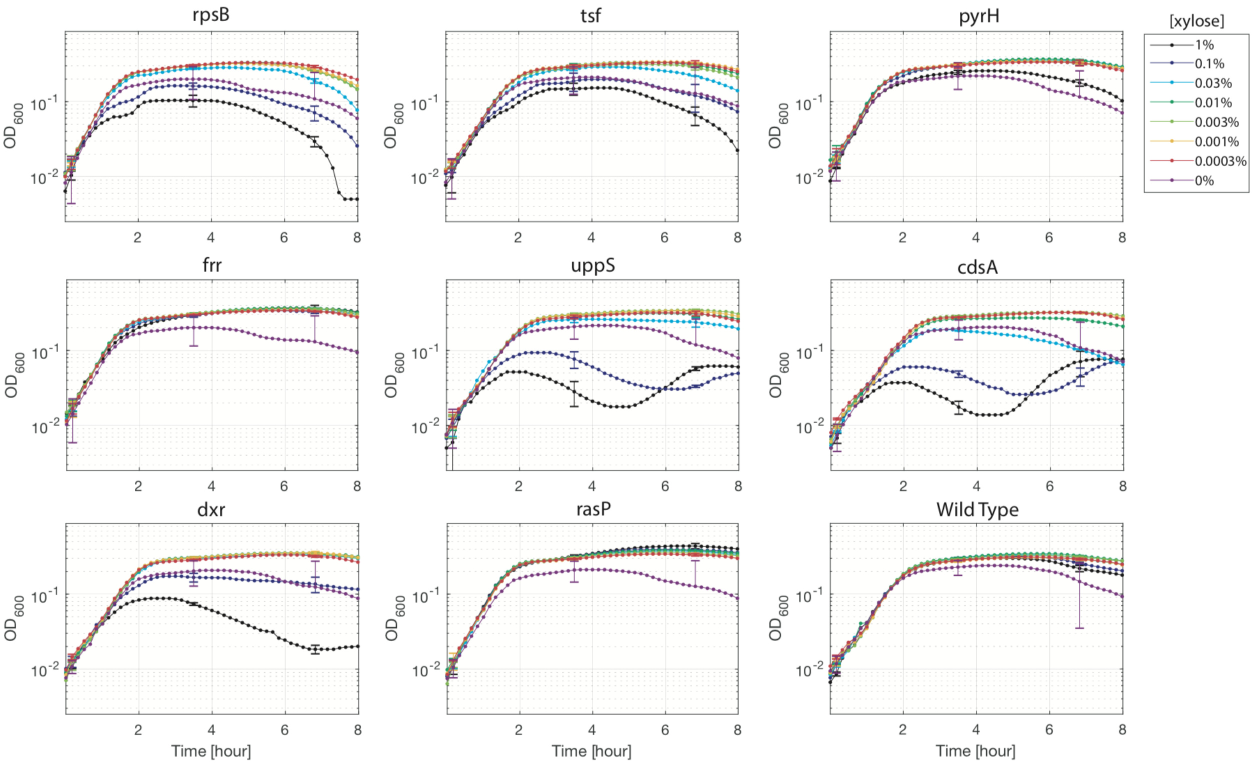
41 - Figure 4.12 - PuppS regulation is transcriptional

RNA was extracted from *B. subtilis* containing the P$_{uppS}$-lux reporter grown in MOPS media with and without amino acid supplementation and subjected to qPCR. Error bars represent the standard error between the replicates. Fold change is relative to the transcript abundances in the 6 amino acid condition. qPCR levels were normalized to the constitutively expressed genes *recA* and *gyrB.*

4.5 Perturbation of gene expression correlation within the cluster

We observed a tight co-regulation of the genes in the TCE cluster, most of which are essential and involved in key growth-related processes. Furthermore, their expression has been modulated and tuned for different species. This suggests that a balance between these processes may be key for efficient growth. Therefore, we investigated if perturbing this stoichiometry between the genes of this cluster would be detrimental to growth.

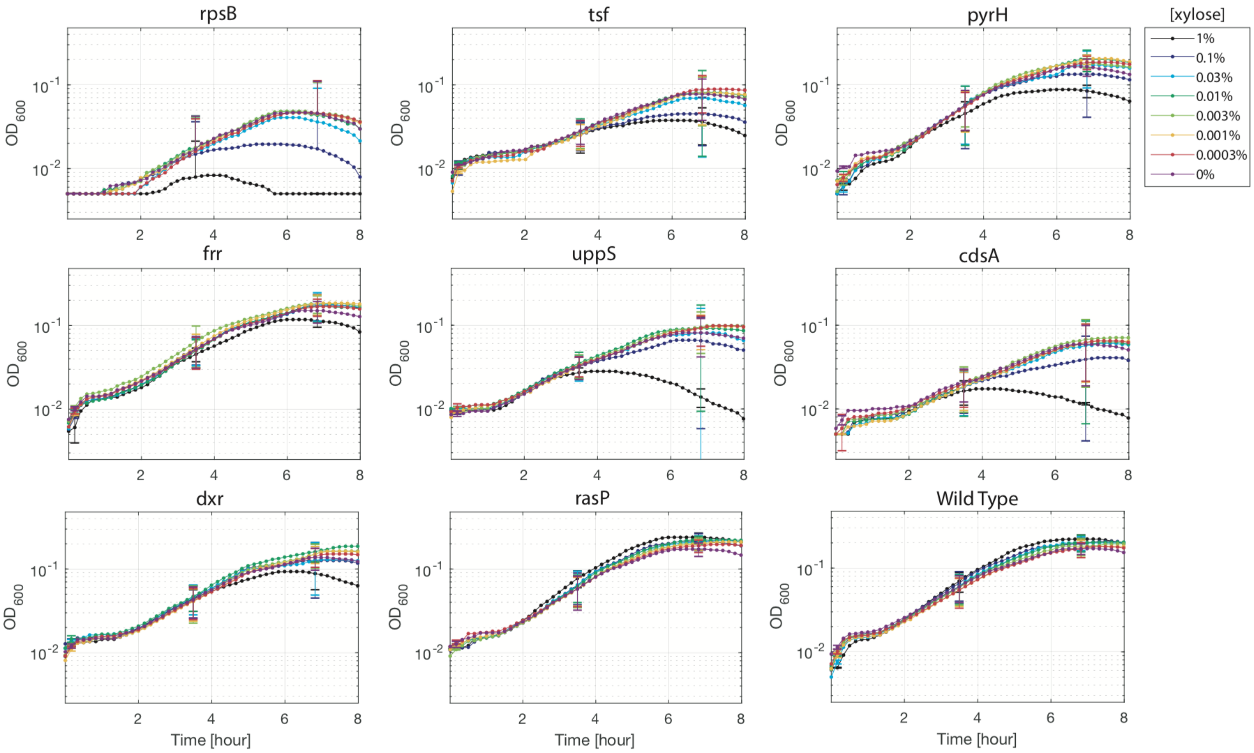
To this end we again used the CRISPRi system. We hypothesized that if balance between these genes is important, this would be most pronounced at fast growth rates where noise would be amplified and there is less time to correct disbalances. We measured the growth of strains knocking down expression of each gene individually in the cluster in both LB and MOPS media, our representatives for fast and slow growth media respectively. Surprisingly we saw that in LB medium, the majority of gene knockdowns did not affect the growth curves except in some cases with highest xylose concentration (1%), i.e. the strongest knockdown condition (Figure 4.13). For example, *rps* and *tsf* which after an early fast growth period had a slower growth rate followed by a lower maximal OD. We believe the initial growth speed occurs due to the time required for the sgRNA to be transcribed and for dilution of older proteins to occur through cell division. *pyrH*, *frr and rasP* knockdowns showed growth curves near-identical to the wild type at all presented levels of knockdown. (Figure 4.13) Upon induction with 0.1% and 1% xylose, *dxr* knockdown caused both cell populations to die (Figure 4.13). In the *uppS* and *cdsA* knockdowns with the same level of transcriptional repression, following the protein dilution that occurs over the first two hours we observed a reduction of cell mass of the culture followed by a regrowth of the culture. As dCas9 blocks transcription by hindering RNA-polymerase progression and we could not predict a transcription start before *cdsA* (Table 2), potentially the growth defects we see from the *uppS* knockdown are derived from a disruption in *cdsA* function as they most likely are on the same transcript.



42 - Figure 4.13 - Effect on gene knockdown in LB

Plate reader measurements of the different CRISPRi knockdown strains grown in LB. Each strain expresses dcas9 and an sgRNA under the control of a xylose inducible promoter which each target a different gene (labeled above each plot). The colour of the bar represents the different xylose concentrations that cells were exposed to at time 0 which are indicated in the legend in the top right of the figure. The error bars represent the standard deviation at two time points.

In comparison, in the slow growth media, MOPS, we observed different responses to gene knockdown, which was not consistent with the LB media results (Figure 4.14). *rasP*, *frr*, and *pyrH* behaved similarly between the two conditions (Figure 4.14 and 4.15). Interestingly the *rpsB* knockdown strain was unable to grow well in MOPS media and began to die during the measurement period even at 0% xylose. *frr* also was unable to reach the same maximum OD in the 0% xylose condition. Both these genes are known to be highly growth rate-sensitive (Borkowski et al. 2016) and therefore in the slow media when expression of these genes are much lower they may be more sensitive to the basal repression caused by CRISPRi. This would suggest that our hypothesis, that synchrony of these genes at fast growth rates is important and controlled by expression of this gene cluster is incorrect and perhaps it is expression at slow growth that needs to be balanced. Alternatively, because both these genes are very highly expressed in the LB condition, the sheer number of DNA-associated RNA polymerases may physically block the dCas9 from interacting with the DNA and lessening the effect of the knockdown. This effect was not observed with *uppS* and *cdsA* which appeared to be less affected by the knockdown, only showing significant growth defects at 1% xylose, and minor effects at 0.1%. Also, strikingly, in MOPS media, the knockdown of *dxr* expression resulted in no change to the growth rate, whereas it was significantly affected at 1% xylose in the LB condition. We compared these growth rates to a wild type and not a strain containing the dCas9 without a guide RNA, however we assume the effects of expressing the dCas9 protein in isolation is minimal in both conditions seeing that knockdown strains such as *rasP* grew identically to the wild type.

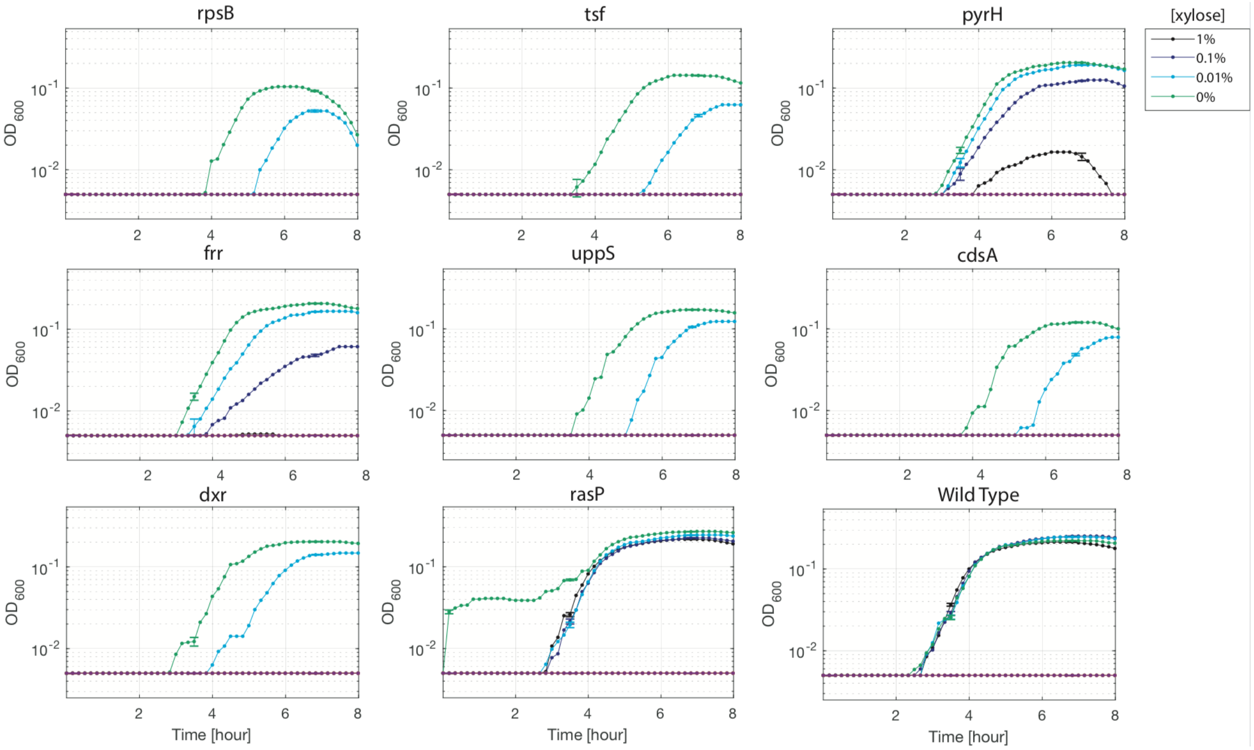


43 - Figure 4.14 - Effect on gene knockdown in MOPS

Plate reader measurements of the different CRISPRi knockdown strains grown in MOPS + glucose. Each strain expresses dcas9 and an sgRNA under the control of a xylose inducible promoter which each target a different gene (labeled above each plot). The colour of the bar represents the different xylose concentrations that cells were exposed to at time 0 which are indicated in the legend in the top right of the figure. The error bars represent the standard deviation at two time points.

Growth is highly sensitive to transcriptional perturbation of TCE cluster genes during transition from lag phase to exponential phase

Growth out of the lag phase has been relatively poorly studied in comparison to the exponential phase, however it has clear implications in the growth of bacteria in natural environments. Therefore, using the same CRISPRi strains as previously, we induced knockdown from inoculation and measured their growth out of the lag phase. We found that in LB many xylose concentrations that were permissible for normal growth in the log phase were lethal to lag phase cultures. All genes, excluding the non-essential *rasP*, displayed a dose-dependent response to xylose-induced knockdown (Figure 4.15). As a different overnight culture is needed for each knockdown, it cannot be guaranteed that the starting cell numbers and time spent in stationary phase is identical between all the samples. Therefore, it is only correct to compare different xylose concentrations in one knockdown strain and not each strain to one another. The greater sensitivity of these genes coming out of lag phase could suggest an important role in regulation of this cluster in the preparation growth in the exponential phase. A key observation here is that whilst the length of lag phase changed as the strength of the knock-down increased, we see some minor long-term effects on the growth rate in exponential phase (noted by a change in slope particularly in the frr knockdown). Based on this preliminary experiment, one could speculate that the early expression of these genes sets the expected growth rate in the media.



44 - Figure 4.15 - Effect on gene knockdown on recovery from stationary phase

Plate reader measurements of the different CRISPRi knockdown strains grown from a LB stationary phase culture transferred into fresh LB media. Each strain expresses dCas9 and a sgRNA targeting a different gene which is under the control of a xylose inducible promoter. The colour of the bar represents the different xylose concentrations that cells were exposed to at time 0 which are indicated in the legend in the top right of the figure.

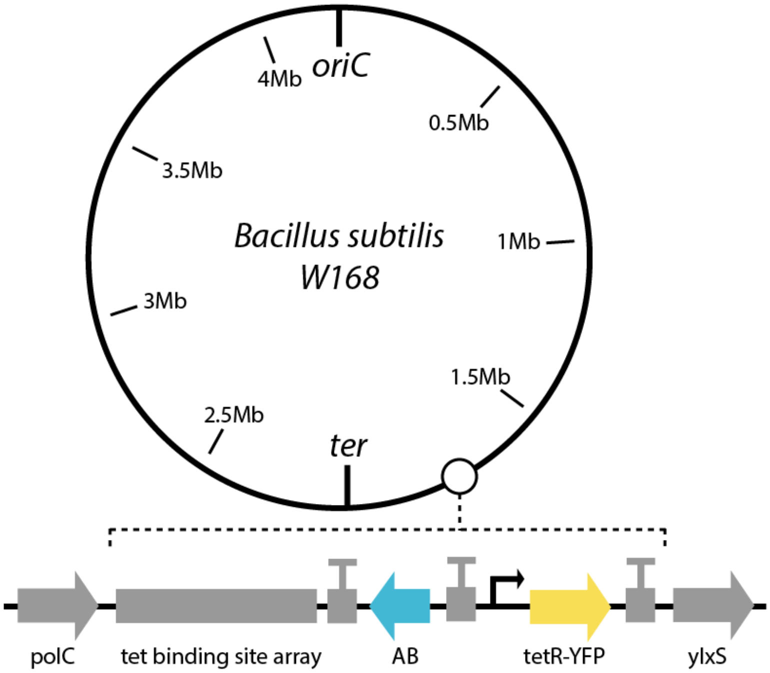
4.6 Role of transertion during expression of the gene cluster

Unlike eukaryotes, bacteria are able to couple translation and transcription as they occur in the same cell compartment. This means that as a gene is being transcribed, ribosomes are already attaching to the nascent transcript to start translation. This for one leads to quicker response of protein production and is especially true on long transcripts such as operons. It is hypothesised that if a gene encodes a membrane-bound product it is also coupled with the insertion of the nascent protein into the membrane in a process referred to as transertion (Matsumoto et al. 2015; Libby, Roggiani, and Goulian 2012). This stipulates that there is a chain between the membrane bound synthesised protein, to the ribosome, to the transcript, to the RNA polymerase to the DNA, which is altogether thought to pull the genomic DNA closer to the membrane (Libby, Roggiani, and Goulian 2012). This has been observed in *E. coli,* where breaking the chain between transcription and membrane insertion, through transcriptional targeting antibiotics, resulted in DNA condensation (Gorle et al. 2017). There are many hypothetical advantages of such a mechanism, including the idea that subsequent membrane-bound proteins require less time until they are positioned in their active site. This is especially true for proteins which form complexes or interact with each other in the membrane when positioned near each other on the chromosome. The force of transertion has also been thought sufficient to shift the nucleoid into an expanded state that allows better access to ribosomes and transcription factors and to assist in chromosome segregation during DNA replication and cell division.

Further investigation into the genes in the TCE cluster reveals that many of them have been shown to be membrane-associated. Interestingly, although *rpsB*, *tsf*, *pyrH* and *frr* do not have annotated membrane binding domains, several experiments have shown that they localized at the cell surface or are membrane-proximal (Hahne et al. 2008; Wilkins, Beighton, and Homer 2003; Lewis, Thaker, and Errington 2000; Gagyi et al. 2004). *cdsA* has been found localised at the septum (Nishibori et al. 2005) and *rasP* has well annotated intra-membrane domains. Only *uppS* and *dxr* lack direct experimental evidence of their cell localization, however the products of their enzymatic activity directly integrate into the membrane, such that localization would facilitate enzyme function. Additionally, we observed that in the *B. subtilis* genome, there exists the flagella gene cluster containing several proteins that form a membrane embedded protein complex upstream of the TCE cluster. Furthermore, we see that the TCE cluster is not conserved in species in which co-transcription and translation are not coupled, such as in Planctomyces (Jogler et al. 2012).

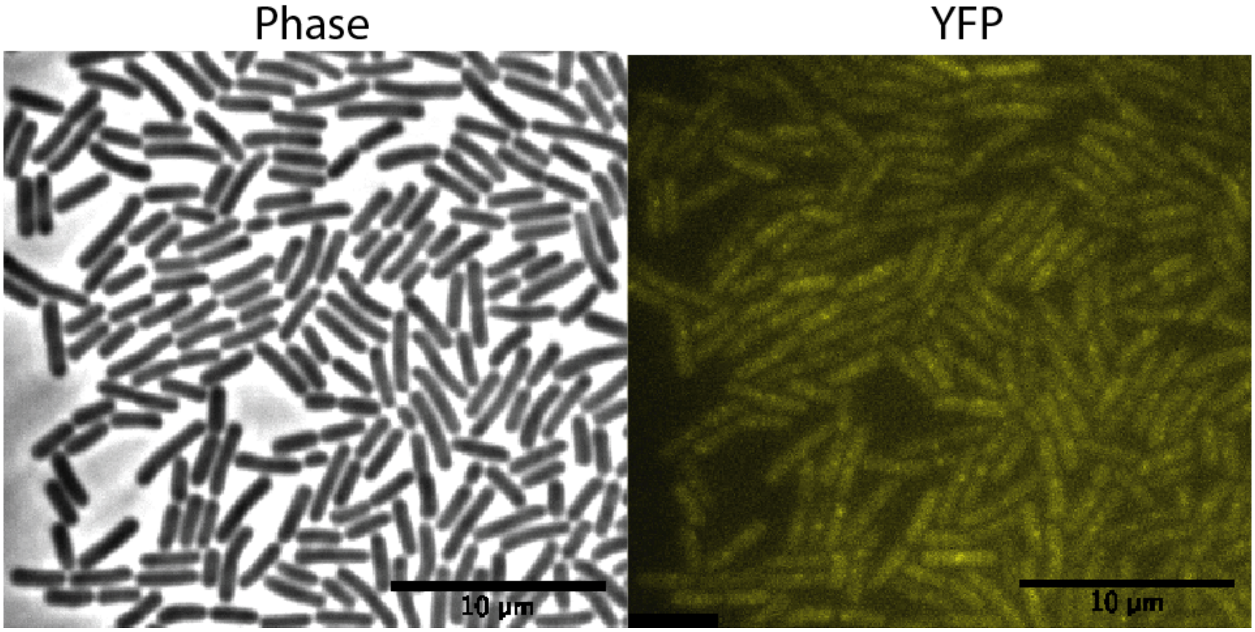
In order to test this hypothesis, we constructed a mutant with a TetR-YFP fusion and inserted a terR binding array downstream of the gene cluster (Figure 4.16). This was placed after *polC* and not directly after the cluster in order to reduce the risk of interfering with any essential processes. We visualized the cells with phase microscopy to observe the position of the TetR binding array in each cell, identified as a focus of light after expression of the TetR-YFP fluorescent protein (Figure 4.17). To measure if this focus was being pulled towards the membrane, we measured the distance of each focus from the longitudinal centre of the cell. Given that we are viewing the foci on a 2D plane whilst the DNA can move in 3 dimensions, one would not expect even in positive cases that foci would only be found distant from the centre. This is because foci could appear in the middle of the cell but still be bordering the cell wall on a vertical axis. However, if a foci would only be found localized to the centre of the cell, there would be very few occurrences of foci at the edges of the cell.

Viewing the foci, however, revealed no bias of the genome to the longitudinal walls of the cell and we saw a mostly even distribution along the cell width (Figure 4.18). This distribution did not match the behaviour of DNA in other hypothesized transertion systems in *E. coli* (Gorle et al. 2017). A control strain would need to be created mainly positioning the TetR binding array next to known membrane proteins or with an inducible membrane protein to know the distribution we should expect in genes we expect to transert and not transert. Rifampicin could then be applied to the control and experimental mutants to break up transertion to determine if that has an effect on the distribution. This was attempted with the *tetB* and *lacY* genes, however, positive clones were not successfully created. Still, the observation that there was an almost uniform distribution of the loci along the cell width suggests that any genome movement towards the cell wall caused by the transcription and translation of the cluster genes is highly unlikely.



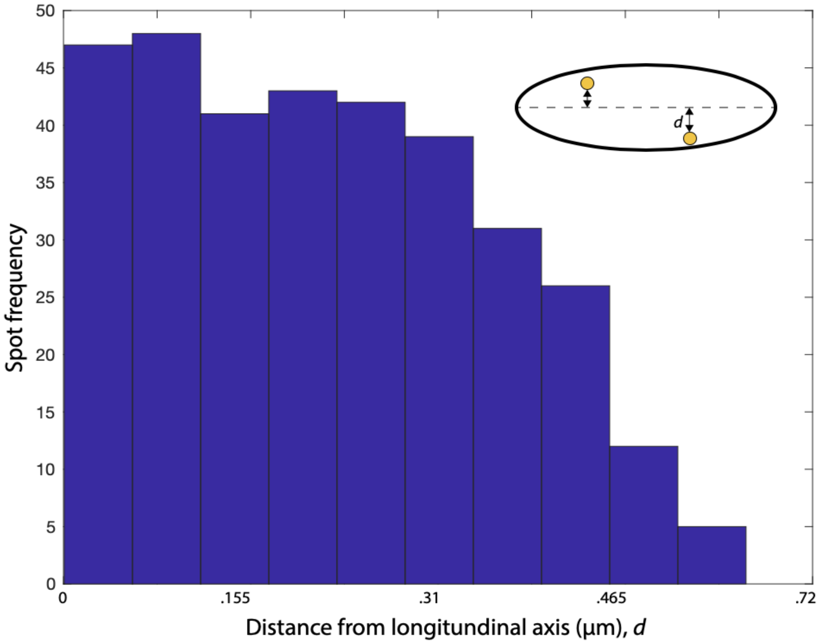
45 - Figure 4.16 - Schematic showing genomic integration to visualize genome localization

The construction of a strain to investigate the transertion potential of the TCE cluster. The area under the dotted line represents the integrated DNA. The insert was placed downstream of the TCE cluster after the polC gene. *polC* and *ylxS* are the endogenous genes and remain unchanged. The 'T' shape represents terminators.



46 - Figure 4.17 - Fluorescent microscopy of the polC-TetR array TetR-YFP mutant

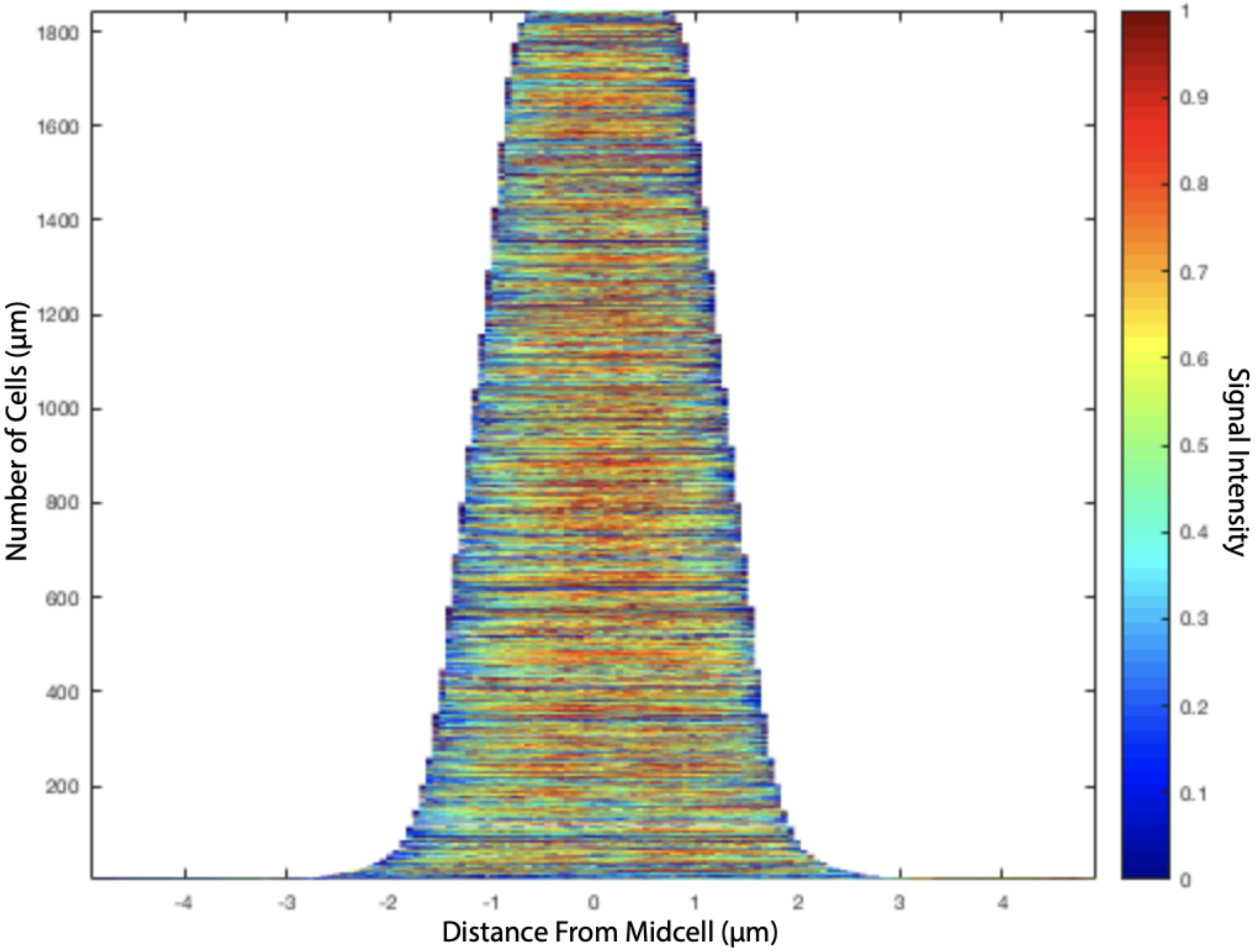
Examples of the phase microscopy images taken of the polC-TetR array TetR-YFP mutant used for foci detection and measurements. Samples were taken from their respective media during exponential phase. Samples were spun down to increase concentration before visualization under the microscope when required. Scale bar in the bottom right corner represents 10µm.



47 - Figure 4.18 - No evidence was found that the gene cluster localizes the DNA to the cell wall

Bacillus subtilis containing a tetR binding array localized downstream of the TCE cluster and a TetR-Yfp fusion protein was grown in LB and viewed under phase microscopy. The genomic loci of the localization cluster in relation to the cell wall was measured. The histogram shows the frequency of distances between the longitudinal centre of the cell and the spot signals (*d*) that were detected.

During bacterial cell replication the key challenge is to correctly segregate the DNA into the two daughter cells. In many cases the origin of the two genomes are pulled to the soon to be old poles and termini are found near the septum (Wheeler and Shapiro 1997). As seen in Chapter 3, the TCE cluster is often localized near the terminus on the genome. *cdsA* has also found to be localized specifically to the membrane at the septum (Nishibori et al. 2005) potentially pulling the terminus also towards the septum through transertion forces. Therefore, we also measured the distance between the foci and the midcell. The foci appeared to be well distributed across the length of the cell and therefore we do not suggest from these data that the cluster is relevant for chromosomal localization during cell division.

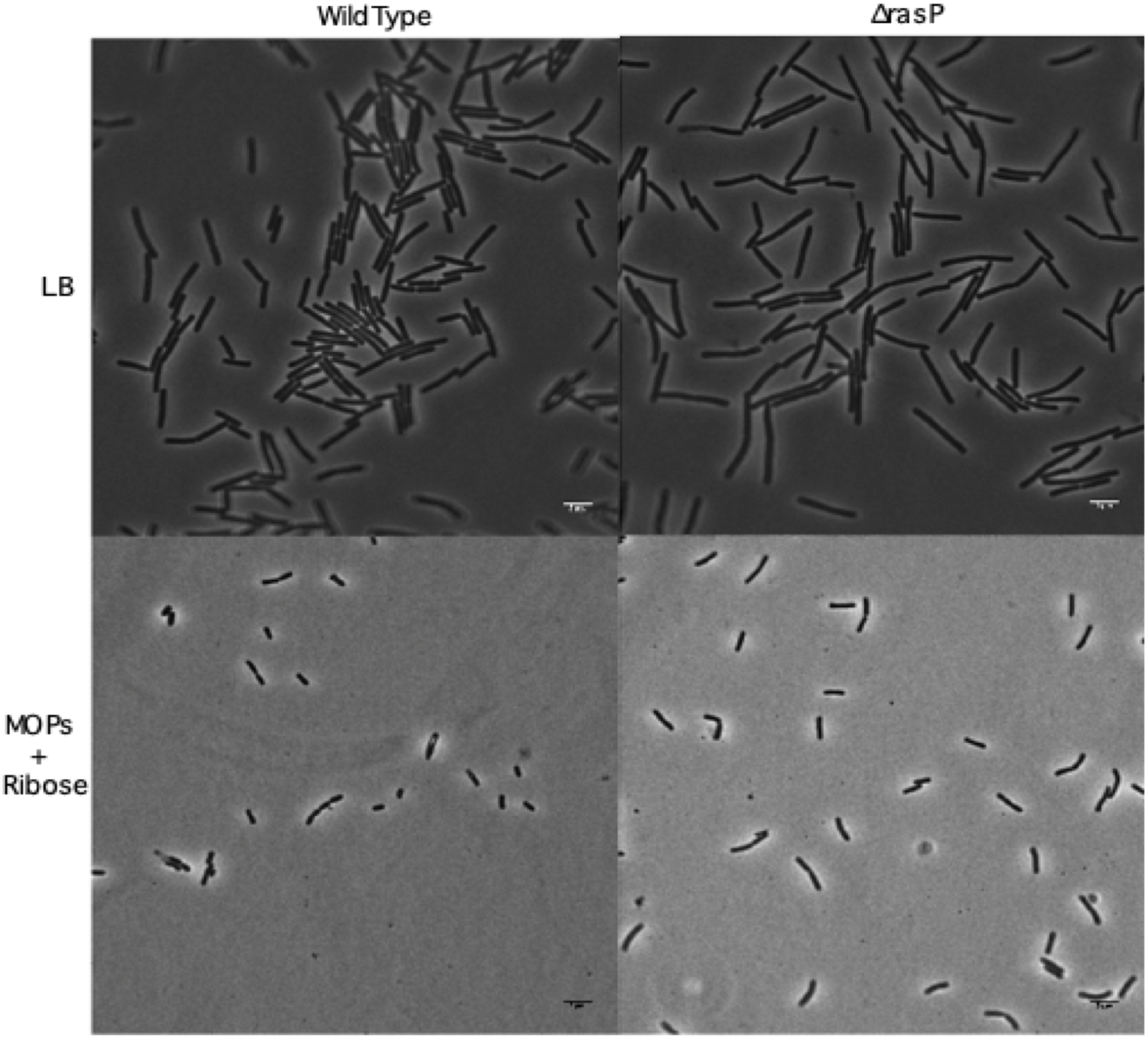


48 - Figure 4.19 - No evidence was found that the gene cluster localizes the DNA to mid cell

The genomic loci of the localization cluster in relation to the midcell was measured. The demograph shows average signal strength along the length of *Bacillus subtilis* containing a tetR binding array localized downstream of the TCE cluster and a TetR-Yfp fusion protein grown in LB and viewed under confocal microscopy. The signal strength of each cell is displayed along its longitudinal axis and they are sorted by the length of the cells with red representing the strongest signal.

4.7 *rasP'*s role in modulating the relationship between cell size and growth rate

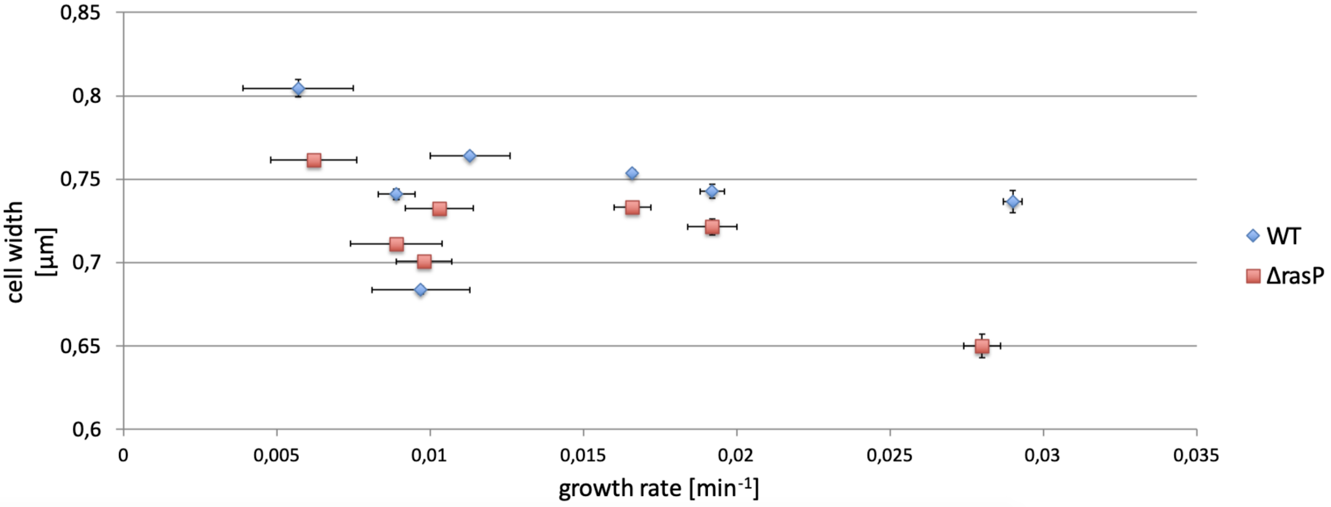
It is still an ongoing question in microbiology how the regulation and mechanism of cell size in bacteria works. As outlined in the introduction, early research in bacterial growth revealed that cell size positively correlated with growth rate (Schaechter, Maaloe, and Kjeldgaard 1958) and later it was elucidated that cell size homeostasis is ensured by the addition of a fixed amount of cell size per round of division (Taheri-Araghi et al. 2015). What is currently missing from the equation is a conclusion to the question, how does the cell translate their growth rate into how much size they should add? Presumably this mechanism needs to integrate a measure of growth rate (for example the expression of ribosomal proteins) with the control of cell envelope and division machinery which are both directly impacted by changes in size. With the TCE cluster containing all three of these elements we hypothesized that these genes may control the relationship between growth rate and size. *rasP* is a unique member of the TCE gene cluster as it is the only non-essential gene and can be deleted without significant changes to the fitness of the cell. Cells which have been depleted of *rasP* have been shown to be shorter without significant changes to the cell's growth rate (Bramkamp et al. 2006). This is predicted to occur as RasP digests FtsL which is a last stage divisome protein. If RasP levels are high, FtsL levels will be low, resulting in a delayed division and cells growing larger, hence in the absence of RasP, FtsL can accumulate quicker, leading to faster (earlier) division and smaller cells. As discussed in the introduction, when bacteria are growing at a faster rate their cell size is also larger at division. We propose a mechanism that by coupling expression of *rasP* with *rpsB* and other translation factors, the cell is able to link the timing of division (and therefore size) with growth rate through the expression of *rasP*. Consequently, if this was the case, by deleting *rasP* we should not only see smaller cells but cells where the relationship between growth rate and cell size is weaker.



49 - Figure 4.20 - rasP has a role regulating growth rate with cell size

The difference in cell size in context of *rasP* depletion was measured. Examples of the phase microscopy images used for cell size measurements are shown. Samples were taken from their respective media during exponential phase. Samples were spun down to increase concentration before visualization under the microscope when required. Scale bar in the bottom right corner represents 5µm.

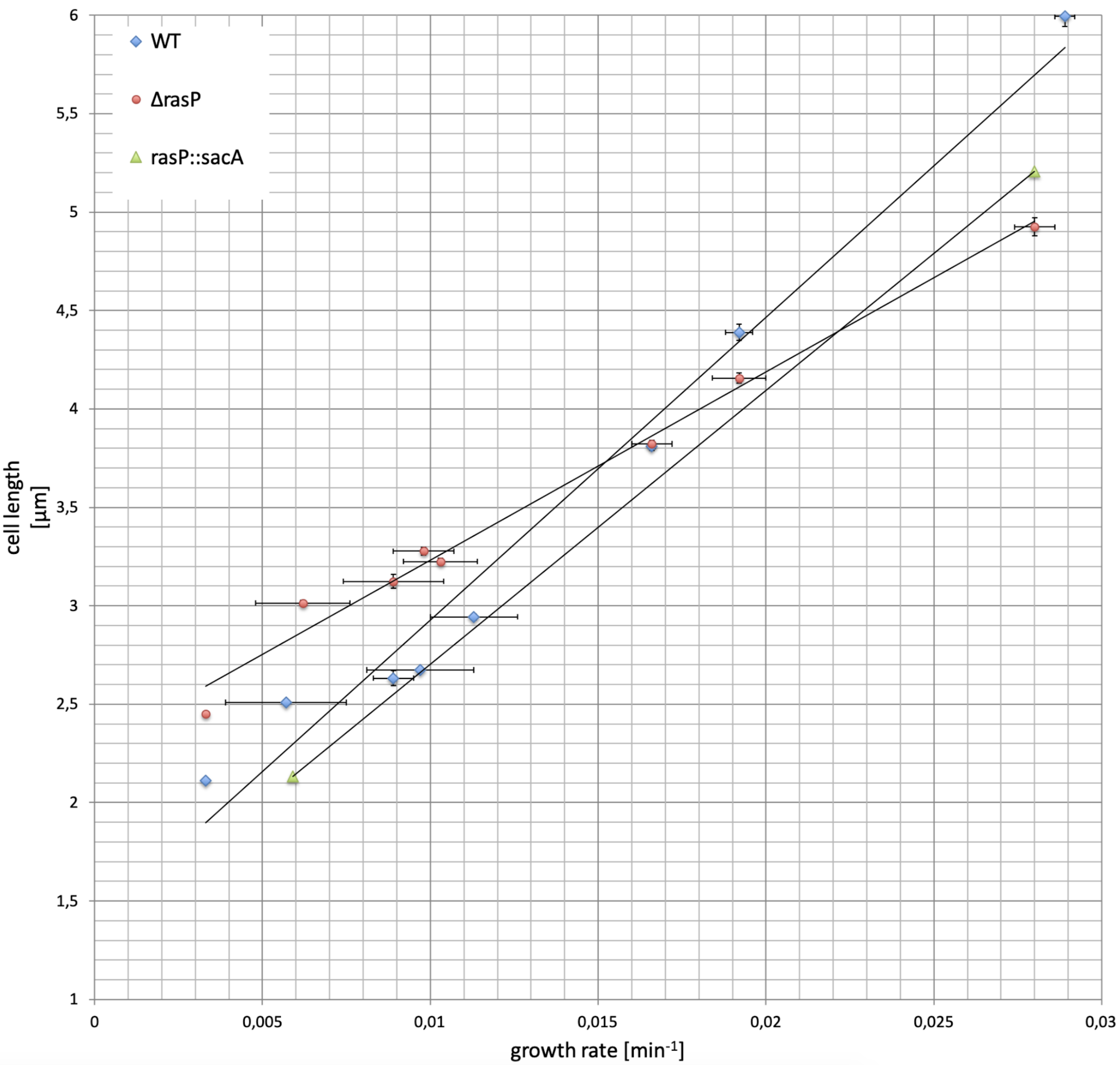
Jesica Bzdok, a master's student in our lab, created a *rasP* deletion mutant and measured the size of the cells at different growth rates. To ensure we observed the full range of the *B. subtilis* growth rate range, the cells were growing in various media with different carbon sources and amino acid supplementation. In order from highest nutritional quality to lowest, LB media, MOPS media + glucose and amino acids (aa), MOPS media + glucose, MOPS media, + glycerol and aa, MOPS media + glycerol, MOPS media + ribose and aa, and MOPS media + ribose. Our fastest growing condition was LB medium, where a doubling time of ~24 minutes was achieved, and the slowest was MOPS media with succinate as the carbon source resulting in a ~210-minute doubling time. The *rasP* mutant did not have significant growth rate differences compared to wild type. *B. subtilis* cells are known not to vary greatly in their width at different growth rates, therefore we rely solely on cell length as a measure of cell size (Sargent 1975). We were able to confirm this in all our growth conditions and in the *rasP* deletion mutant (Figure 4.21).



50 - Figure 4.21 - Deletion of rasP does not change cell width

The effect of *rasP* deletions on cell width in *Bacillus subtilis* at different growth rates. Each data point represents cells from different strains and media. In order from highest nutritional quality media to lowest, LB media, MOPS media + glucose and amino acids (aa), MOPS media + glucose, MOPS media, + glycerol and aa, MOPS media + glycerol, MOPS media + ribose and aa, and MOPS media + ribose. Vertical error bars represent the standard error in the cell width, horizontal error bars represent the standard error in the growth rate.

As measured previously by Bramkamp et al, we observed the shorter cell phenotype in LB when *rasP* was deleted. However, this observation did not hold in our slower growth conditions where in fact we saw the *rasP* mutants were larger (Figure 4.20). Across wild type samples we measured a change in average size from 6.004 µm (±0.053SE) in the fast growing LB media to 2.509 µm (±0.012SE) in the slow MOPS media (Figure 4.22) matching the size:growth rate relationship previously measured by Taheri-Araghi et al. The mutant had average sizes between 4.92 µm and 2.45 µm (Figure 4.22). The linear relationship between length and growth rate was different between the wild type and the mutant and as we predicted, we saw a decrease in the relationship (slope) in the mutant with 1.594 µm/λ[hour] compared to 2.570 µm/λ[hour]. Despite the range of different carbon sources and nutrient content, all 8 media correlated in a linear relationship for both the wild type ($R^2 = 0.9792$) and the *rasP* mutant ($R^2 = 0.9878$). By complementing *rasP* in another genomic context, we were able to mostly restore the wild type relationship between growth rate and cell size (2.268 µm/λ[hour]), however, the cells were not as large as the wild type but still bigger than the complete deletion (Figure 4.22). It should be noted that the *rasP* complement was expressed under a constitutive promoter (P_{liaG}), which was not modulated at different growth rates, therefore acting similarly to a constitutive promoter in this instance and the overexpression of the *rasP* complement could explain smaller sizes overall. Further experiments are needed where we express *rasP* at different levels. This would help us decipher *rasPs* role better in cell size determination. As we were able to restore the balance between cell length and growth rate with a *rasP* complement at a different genomic locus, we suggest that this function of *rasP* is independent of genomic localization.



51 - Figure 4.22 - rasP knockout mutants are smaller than wild type in rich media but smaller in poor media

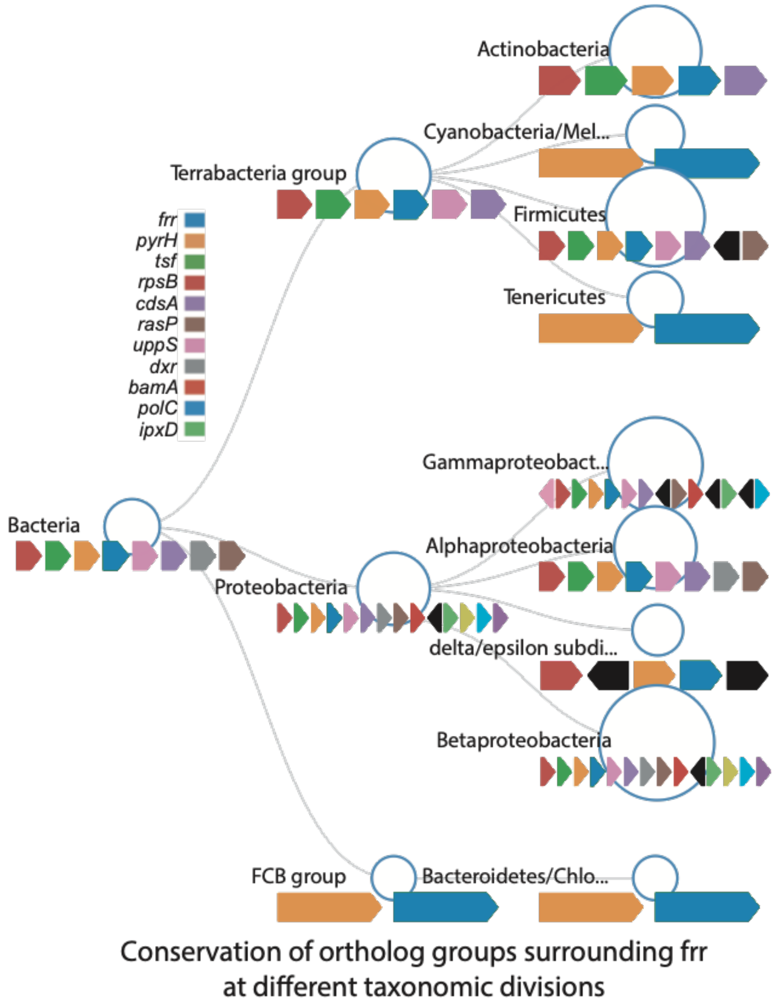
The effect of *rasP* mutations on cell length in *Bacillus subtilis* at different growth rates. Each data point represents cells from different strains and media. Vertical error bars represent the standard error in the cell length, horizontal error bars represent the standard error in the growth rate. Linear regressions were fit to the data points each strain type. rasP was complementation in the sacA locus under the control xylose inducible promoter (1% xylose). In order from highest nutritional quality to lowest, LB media, MOPS media + glucose and amino acids (aa), MOPS media + glucose, MOPS media, + glycerol and aa, MOPS media + glycerol, MOPS media + ribose and aa, and MOPS media + ribose.

FtsL is not the only target of RasP as it also cuts RsiV and RsiW, which are anti sigma factors to their respective sigma factors σ^V and σ^W and block their activities until they are proteolysed (Zweers et al. 2012). σ^W is responsible for regulating genes involved in cell wall homeostasis

(Eiamphungporn and Helmann 2008) and σ^V is responsible upregulating genes involved in lytic resistance (Zellmeier et al. 2005). It is possible that in the absence of RasP there is a lack of σ^V and σ^W activity as a result of their respective anti-sigma factors not being proteolysed. Therefore, to test this scenario we did preliminary measurements of the cell size of mutants deleted in σ^V and *sig*^{W} separately (Figure 8.1). We found that the sizes of the σ^W mutant matched the length and slope of the *rasP* complement mutant. The deletion of σ^W resulted in similar changes as seen in the rasP deletion mutant, however not as dramatic (a slope of 2.04 µm/λ[hour]). Interestingly, the intersect between the slopes of the WT, *rasP* deletion mutant and the σ^W deletion mutant meet at a doubling time of 50 minutes. With these data it is presently not possible to deduce whether the effect we see in the rasP mutant is partially due to its role in σ^W activity or if the effect of the σ^W depletion is independent as we do not quantitatively know the impact of RasP on σW *in vivo*.

4.8 Phylogenetic Analysis of the TCE cluster

Unlike previously discovered gene clusters, the TCE consists of genes which are mostly from as of yet unrelated and diverse physiological processes (Nikolaichik and Donachie 2000; Tamames et al. 2001) and there is very little literature detailing interactions between these proteins. Therefore, this cluster stands out as highly unique even amongst the already known highly conserved gene clusters. Rare for a cluster of its size, the TCE cluster is found widely conserved across the bacterial kingdom including the phyla of Proteobacteria, Actinobacteria and Firmicutes with neighbourhood conservation scores surrounding frr being 24.91, 31.58, and 33.99 respectively (Figure 4.23). Here a neighbourhood conservation score is the average conservation of the top 50 most conserved genes in the neighbourhood. The maximum score is 50 representing 100% conservation of the top 50 most conserved neighbours. In the phylum Bacteroidetes the cluster was not conserved and had a conservation score of 15.22.



52 - Figure 4.23 - Cluster is broadly dispersed across the bacterial kingdom

The conservation of the translation-cell envelope cluster in different Bacteria taxonomic groups. The genomic neighbourhood is centred around *frr* and overlaid on a taxonomic tree (NCBI definitions). The size of the circle at each taxonomic level represents the conservation score. The arrows below the circles represent the most conserved genes (if above the significance threshold) at each position surrounding *frr*.

In the gram-negative phylum Proteobacteria we see an extension to the 3’ end of the cluster with several genes involved in outer-membrane maintenance, a physiological feature not present in the other two gram-positive phyla (Figure 4.23). The first is BamA which is important for the assembly and insertion of beta-barrel proteins in the outer membrane. There are also three genes involved in lipid A biosynthesis (*ipxABD*), which is a phosphorylated glycolipid that anchors the lipopalysaccharide to the outer membrane. The last proteobacteria-specific gene associated with the cluster is *rnhB* (Ribonuclease HII) which degrades RNA of RNA-DNA hybrids. The *in vivo* role of this protein is still unknown, however, early experiments have shown that strains lacking these ribonucleases show a temperature-sensitive growth phenotype in *E. coli (Ohtani et al. 2000)* whereas in *B. subtilis* a similar knockout is lethal (Itaya et al. 1999), highlighting an important role in growth. The mutation studies suggest that *rnhB* is more important in gram-positive bacteria such as *B. subtilis,* however, as it is not co-localized with the cluster in these genomes it is less likely that this importance is relevant in context of the gene cluster.

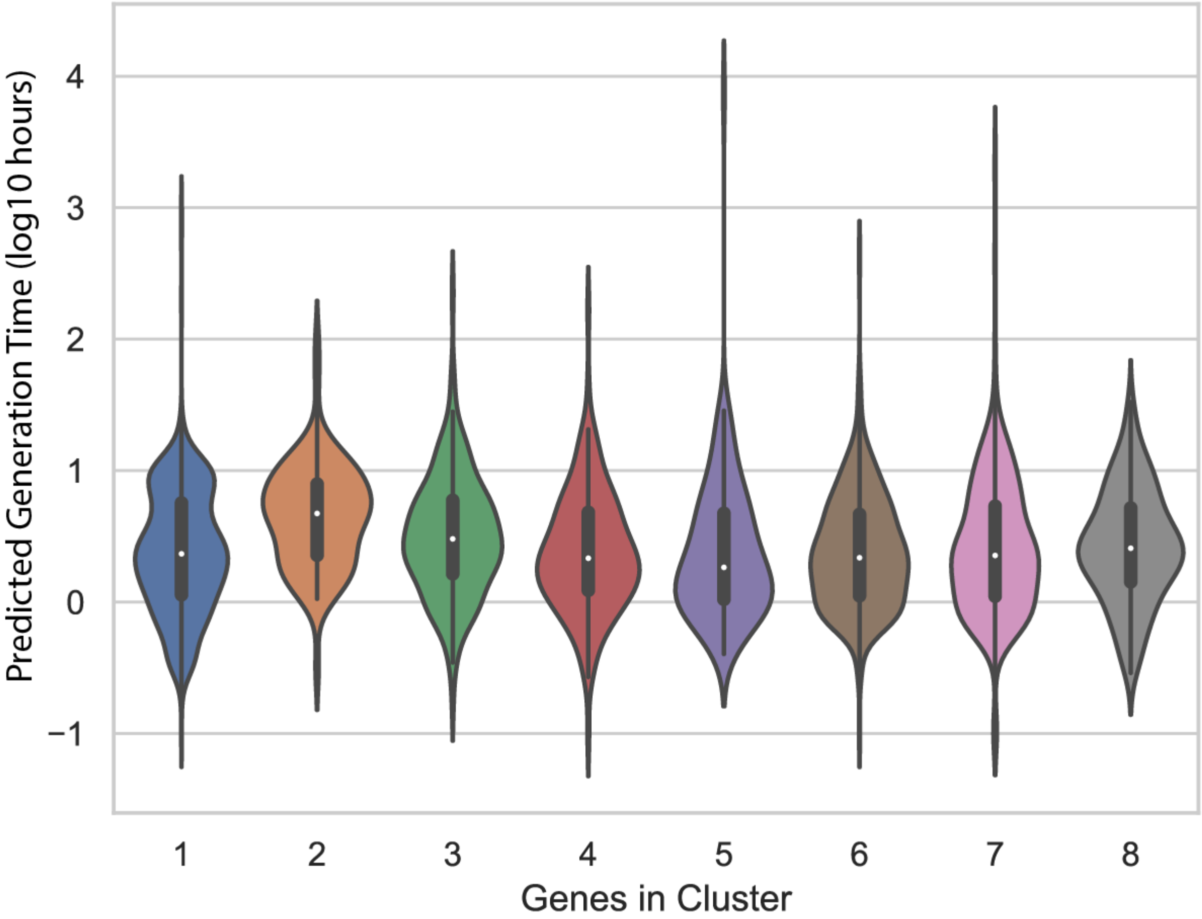
As in Proteobacteria there is also a large 3’ extension to the cluster in Firmicutes, with the genes slowly decreasing in conservation with the length of the cluster (Figure 4.23). Notably several more translation associated genes are found here including *proS*, *rpl7ae*, *infB*, *rbfA*, *truB*, *rimP*, and *rpsO*. Additionally, *polC*, and *nusA* may be found downstream. *polC* is involved in DNA replication and part of the replisome (Hernández-Tamayo et al. 2019) and *nusA* is involved in transcription termination (Mondal et al. 2016). As we mentioned in 4.7, a large flagellar gene cluster lies upstream in many Firmicutes species.

In Actinobacteria, the cluster sees the loss of three genes, *uppS*, *dxr* and *rasP* and the addition of 23 rRNA methyltransferase *rlmN* (Figure 4.23). The *dxr* and *rasP* genes remain tightly conserved together in these genomes and are associated with *ispG* - another gene involved in isoprenoid biosynthesis directly downstream from *dxr*. In Actinobacteria there are two copies of *uppS*, suggesting a duplication event early in the phylum's evolutionary history. Given that *uppS* is normally located between *frr* and *cdsA,* which both remain in the extant cluster, a very localized rearrangement event must have occurred to maintain *cdsA's* status in the cluster, further highlighting the fitness benefits the co-localization provides. Upstream of the cluster (not downstream as in Proteobacteria), *rnhB* can also be found in many Actinobacteria, suggesting convergent evolution in the localization of this gene with this cluster.

Notably the cluster is not heavily present in Bacteroidetes however all the genes remained conserved, with identifiable orthologs being present in over 90% of the Bacteroidetes genomes. Despite the degradation of the cluster in Bacteroidetes many of the individual genes remain co-localized in pairs, namely *rpsB-tsf* (+2 other ribosomal proteins), *pyrH-frr* and *dxr-rasP*. *cdsA* is found mostly without a conserved neighbourhood with minor co-localization with a ribosomal silencing factor (*rsfS*), *ftsH*, and phosphatidylserine decarboxylase (*psd*). Bacteroidetes are widespread and found in varying ecosystems. They are anaerobic, well known for their polymer-decomposing capabilities, and are primarily located in the gastrointestinal tract of animals (Smith et al., 2006; Ley et al., 2009; Thomas et al., 2011). They are gram-negative, chemo-organotrophic, rod shaped and do not form endospores (Woese, 1987; Paster et al., 1994). At this stage, a unique phenotype of the Bacteroidetes, which explains how Bacteroidetes differ from the cluster-containing phyla that might indicate a potential role of the cluster, has not been identified.

Conservation of the cluster is not associated with codon-bias towards fast growth

Only by accurately synchronizing expansion of cell volume and surface is a cell able to achieve a maximum growth speed in a specific media. Therefore, if the TCE facilitated this synchronization we would expect species containing the cluster to also grow faster. Currently growth rate data, especially data in media designed to optimize growth rates, are not available for the breadth of species we have in our dataset to test this hypothesis. We took advantage of the observed relationship between the maximal growth rate and codon bias seen in bacteria (Vieira-Silva and Rocha 2010). This method exploits the varying sizes of tRNA pools in the cell, which shrink (in proportion to the number of ribosomes) and become limiting at fast growth rates (Dong, Nilsson, and Kurland 1996) resulting in preferences for certain synonymous codons over others. This is especially important in highly translated genes such as ribosomal proteins. Thus, cells optimizing for growth rate should have a bias towards the favoured codons in these highly translated genes. Using this method, we calculated the predicted maximum growth rate for each genome in our dataset and compared that size of the TCE cluster (the highest number of cluster genes found in context with one another). Surprisingly, we observed that there was no correlation between these two measurements (Figure 4.24). This suggests that the evolutionary forces acting on fast growth rates and the maintenance of the cluster are independent of each other.



53 - Figure 4.24 - Cluster conservation and growth-rate-codon-optimization is not correlated
Violin plot showing the predicted generation time based on codon bias in highly translated genes in genomes containing different numbers of genes in the TCE cluster. The white circle in each violin represents the median, the thick line extends to the upper and lower quartiles and the width represents the frequency of the data points.

Correlated presence and absence of genes with the TCE cluster

To analyse bioinformatically what is different between species with and without the genes clustered we looked for genes/ortholog groups that are present only in species with the cluster together and vice versa. We categorized each genome cluster as either containing the cluster or not (less than half of the genes are found together) and then for every ortholog group measured how often it appeared in a genome with and without the cluster. Genes were then ranked by the product of their frequency in cluster-containing genomes and their absence in cluster-lacking genomes. This was repeated where genes were ranked based on their frequency in cluster-lacking genomes and absence in cluster-containing genomes. This strategy prioritizes both genes which are highly conserved, and therefore more likely to have a physiological impact, and those which actively evolutionarily correlate with the presence/absence of the cluster. This is contrasted to taking a sum instead of the product of the frequencies. For example, a gene which is poorly conserved and found in only 10% of genomes which coincidentally are cluster-containing genomes would in this case receive a high rank. For similar reasons we did not apply a statistical test such as a Fisher exact test which would more likely identify genes based on their taxonomic distribution.

Genes correlating with the presence of the TCE Cluster	Genes correlating with the absence of the TCE Cluster
Similarity:Contains HPr domain.	Endonuclease MutS2
LexA repressor	ribosomal protein L33
ribonuclease PH	arginine decarboxylase
D-alanyl-D-alanine carboxypeptidase	Transmembrane:Helical
Similarity:Contains HTH gntR-type DNA-binding domain.	hypothetical protein
cell division protein FtsQ	Ribonuclease Y
transcriptional regulator NrdR	Similarity:Belongs to the complex I subunit 6 family.
segregation and condensation protein A	uridine kinase
phosphoenolpyruvate-protein phosphotransferase	Alpha-D-phosphohexomutase alpha/beta/alpha domain I
DNA-directed RNA polymerase omega subunit	4-alpha-glucanotransferase
RNA-binding protein Hfq	Transmembrane:Helical
Similarity:Contains 1 HTH iclR-type DNA-binding domain.	DNA polymerase III delta subunit
Similarity:Contains HTH gntR-type DNA-binding domain.	purine nucleoside phosphorylase
RapZ-like family	HD domain protein
Transport permease protein	hypothetical protein
Transcriptional regulator MraZ	hypothetical protein
uroporphyrin-III C-methyltransferase	hypothetical protein
Similarity:Contains 1 CSD (cold-shock) domain.	GDP-L-fucose synthase
Glutaredoxin	Transmembrane:Helical
Transmembrane:Helical	Transmembrane:Helical

5 - Table 4.3 - Presence and absence gene correlation with the cluster

This table represents the top 20 genes which had the highest correlations of being either absent or present alongside the cluster, sorted in descending order. Ranking is based on the product of the number of genomes a gene appears within a conserved TCE cluster and the number where the gene is absent when the TCE cluster is not conserved (and vice versa for absence).

The ortholog group most correlated with conservation of the cluster is a group of genes which have similarity to HPr domain containing proteins (Table 4.3). HPr proteins are phosphotransferases involved in PTS-dependent sugar transport and carbon catabolite repression. This hints that potentially the cluster may be linked with central carbon metabolism; however, further experiments would be needed to clarify this connection. Conversely the gene correlated with a break-up of the TCE cluster is endonuclease MutS2 which is involved in suppressing homologous recombination and repairing DNA after oxidative DNA damage (Table 4.3). This is surprising as it suggests that after genome rearrangements (potentially from homologous recombination) resulting in the separation of the gene cluster, genes are selectively brought in to reduce further arrangements and locking in the dispersed cluster. We see that arginine decarboxylase and uridine kinase are brought in in the absence of the cluster (Table 4.3). The former is involved in the catabolism of arginine for varying purposes (nitrogen source, ATP) and uridine kinase acts as a salvaging pathway for more

UMP. Both these metabolic processes are close together in the bacterial metabolic network, hinting at potential interactions genes in the TCE may have with these metabolites. There are several poorly annotated genes which are enriched in genomes where the cluster is not co-localised (Table 4.3). One likely reason is that many of the genomes where the cluster is not conserved are less researched and therefore more likely to have novel and under-researched genes.

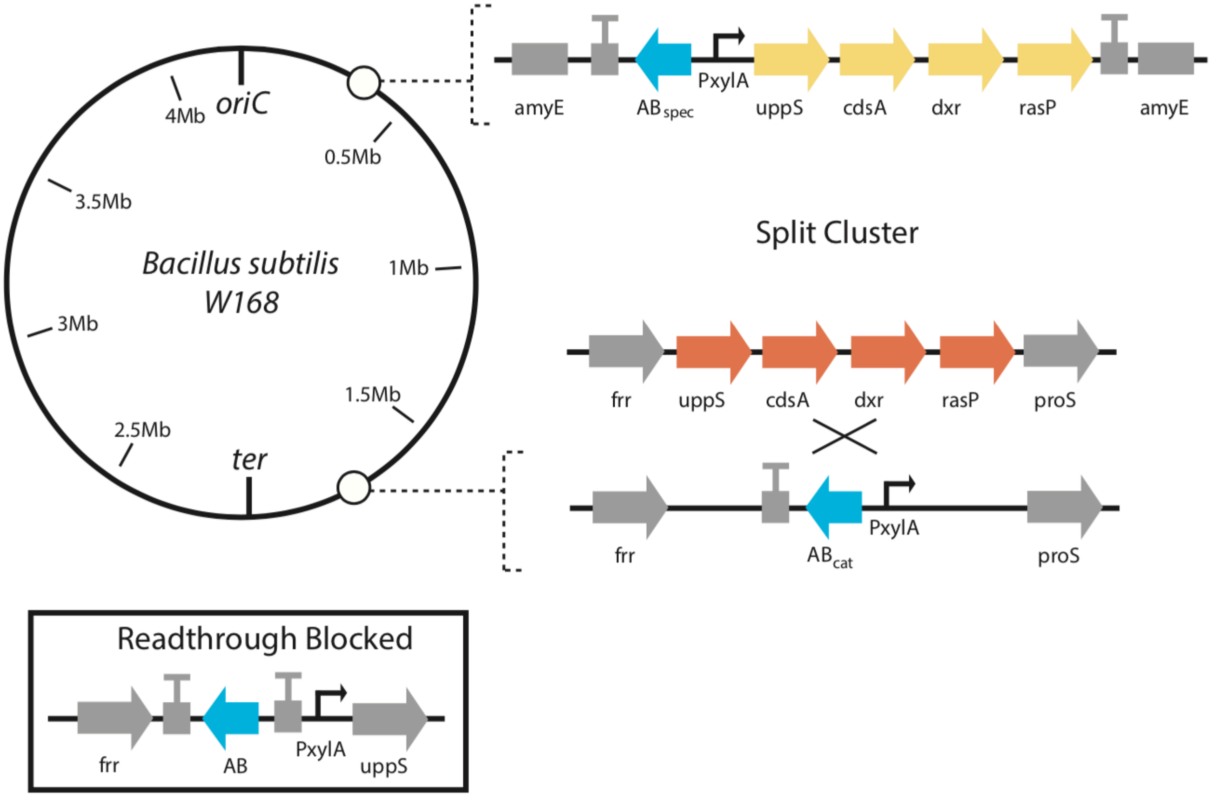
4.9 Delocalization of the TCE cluster

The gene presence/absence correlation analysis in the previous section revealed to us several genes which correlated with the splitting of the TCE cluster. Two of these were arginine decarboxylase and uridine kinase, standing out due to their role in the pyrimidine/arginine biosynthesis pathway which many genes in the cluster also share. *pyrH* is directly required for UDP synthesis. *uppS*, *cdsA* and *dxr* all use CTP or UDP directly or in their downstream substrates, therefore UDP concentrations are highly relevant for their activity. We thus hypothesized the close proximity of the cell envelope genes to the pyrimidine producing enzyme is a form of genomic challenging (Mingorance, Tamames, and Vicente 2004). Local transcription and subsequently coupled translation results in higher local concentrations of the pyrimidine producer (*pyrH)* and the pyrimidine consumers.

To see if the co-localization of the pyrimidine producer with the pyrimidine consumers is important, we attempted to break the association between these genes. We first attempted to insert a large sequence of inactive DNA (40kb) between *frr* and *pyrH* thereby separating the two halves of the cluster. We used inactive DNA that was constructed to minimize transcriptional activity (Zumkeller, Schindler, and Waldminghaus 2018). Whilst we were able to create a vector for the transformation, we were unable to transform it into *B. subtilis*. This may be due to either the large insert size decreasing the transformation efficiency or the insertion of the spacer DNA into the cluster being lethal. Given that the alternative strategies did function the issue was most likely the former. With the propensity of *B. subtilis* to recombine homologous DNA and the spacer DNA being highly similar, it is also likely the inserted sequence would not have been stable and shrunk over time.

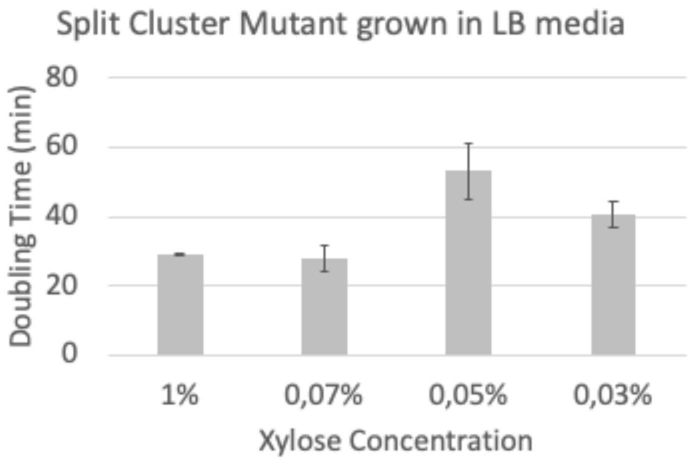
Creation of a TCE split cluster mutant

We next attempted to directly move the latter four genes of the cluster to another location on the genome. First, we transformed copies of the four genes from *uppS* to *rasP* into the *amyE* locus under a xylose inducible promoter with no inducer to reduce the risk of lethality from overexpressing these genes (Figure 4.25). The *amyE* locus is an established integration site for *B. subtilis,* which has been shown not to have different levels of activity compared to other integration sites and does not appear to cause collateral effects to neighbouring genes (Kim, Mogk, and Schumann 1996). We then deleted the original locus of four genes in the presence of xylose (Figure 4.25). To ensure the expression of the genes downstream from the cluster was maintained, the same xylose inducible promoter was inserted before them (Figure 4.25). From the transcriptome data (Figure 4.25) we observed that the expression of the latter four genes in the cluster and their downstream neighbours were all expressed at the same or similar level, which is why we deemed utilizing the same inducible promoter to ensure equivalent expression was closest to the wild type condition. We tested growth of this mutant in various concentrations of xylose and found that the dynamic range of growth rate fell between 0.3% and 0.7% (Figure 4.26). As we were able to get reproducible growth and to ensure sufficient xylose remained in the media over long incubation times and high optical densities we used 1% xylose as our standard concentration (Figure 4.26), which is close to the saturating concentration of the promoter (Radeck et al. 2013). We measured *uppS* expression with qPCR in the wild type and the mutant at 1% xylose in MOPS media + Glucose minimal media and observed no significant difference in expression (Figure 4.28).



54 - Figure 4.25 - Schematic showing how the cluster was split

Construction of strain where the TCE cluster has been delocalized (split cluster strain). The areas shown by the dotted line represent the integrated DNA and current state of the strain's genome. The genes in red were replaced with an antibiotic cassette (blue) through homologous recombination. The genes in grey represent the endogenous genome. AmyE was an integration site and divided in two with the insertion of the second half of the cluster (yellow). The 'T' shape represents terminators. The readthrough blocked strain is also displayed in a separate box. Here, an antibiotic resistance cassette was placed in the reverse orientation between *frr* and *uppS* to ensure that no transcriptional readthrough could occur.

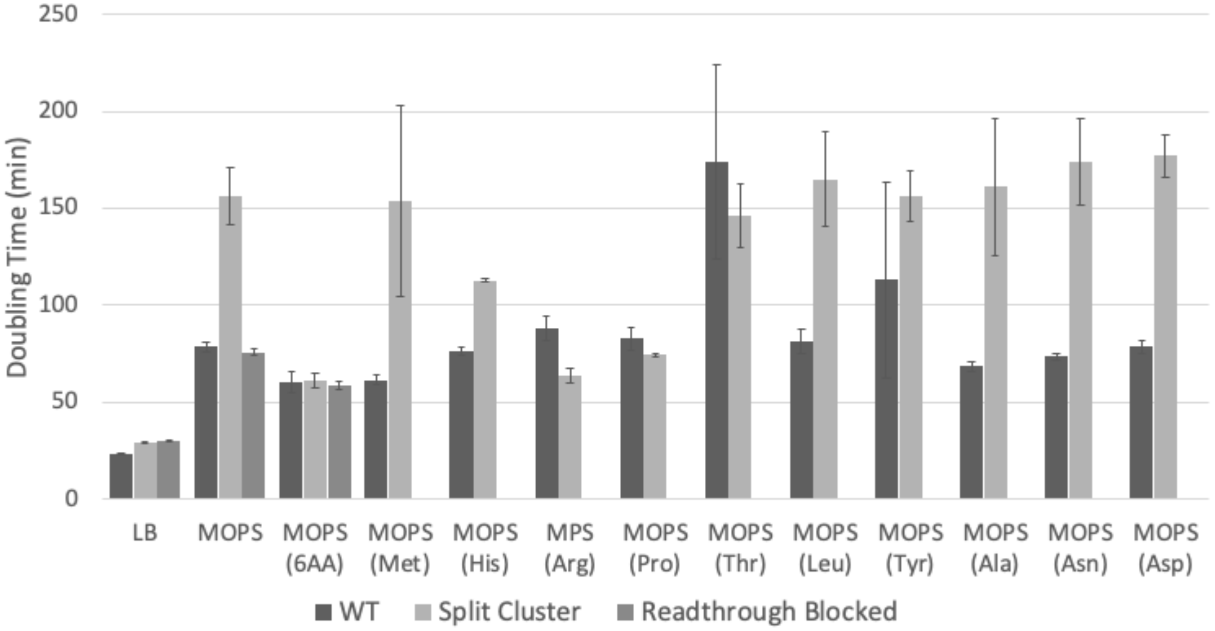


55 - Figure 4.26 - Xylose dependency of the split cluster mutant

Doubling time measurements were taken from plate reader measurements of exponentially growing split cluster mutants grown in LB at varying xylose concentrations. Error bars represent standard deviation between the replicates (n=3).

Splitting the cluster results in arginine auxotrophic behavior

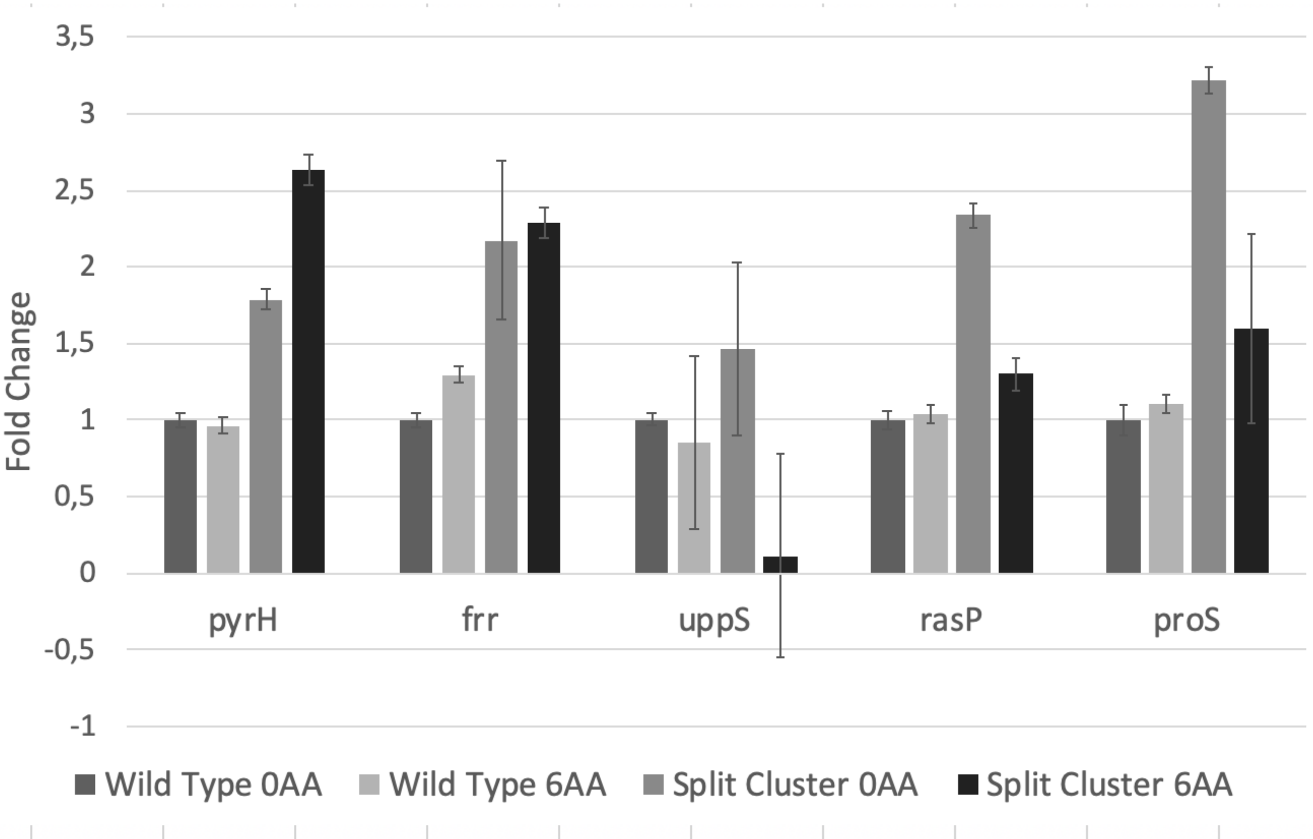
With the completion of our experimental strain we could now examine how the delocalization of the cluster affects the physiology of a cell. First, we wanted to explore the role of pyrimidines in the mutant's metabolism. If pyrimidine and subsequently arginine metabolism has been disrupted by the delocalization of the cluster, we would expect the mutant to grow poorly in media lacking a pyrimidine source. *B. subtilis* is unable to import the majority of pyrimidine molecules, specifically the molecules after *pyrH* in the pathway, UDP and UTP which is where we expect the major dysregulation to occur. Therefore, we decided to test the growth of the split cluster mutant in the presence of different amino acids from the arginine biosynthesis metabolism which shares a common metabolite with the *de novo* pyrimidine pathway, carbamoyl phosphate. We observed that in both LB media and in a minimal media supplemented with multiple amino acids (including arginine) there were no obvious differences in the growth rate between the wild type and the split cluster (Figure 4.27). However, when we removed amino acid supplementation from the media, we saw over the course of many experiments that cells which were exposed to amino acid lacking conditions for long periods of time either regrew later at the wild type growth rate or died. Therefore, to observe how the cells behave after shifting to the media, we repeated our plate reader experiments by pre-culturing cells with amino acids, washing them and placing them in the amino acid deprived media at time zero. Doing this we noted an instant 2-fold increase in the doubling time of the cell from ~78 minutes in the wild type to ~156 minutes in the mutant (Figure 4.27). This phenotype was able to be recovered by supplementing only amino acids related to arginine biosynthesis, namely, arginine, proline, glutamate and glutamine, however, none of the other tested amino acids could restore the phenotype (Figure 4.27). As *proS* is downstream of the cluster, now controlled by the xylose inducible promoter, an explanation for the proline auxotrophic behaviour could be the dysregulation of this gene. However, the *rasP* deletion mutant we created under section 4.7 also had *proS* under the same promoter with the identical inducer levels and we did not observe growth rate defects in this same media and therefore deem this possibility unlikely (Figure 4.22). To confirm the effects, we saw were not due to our regulation strategy using the xylose promoter, we created another mutant where we put the native *uppS* under the control of the xylose promoter, placing an antibiotic resistance after *frr* as in the split cluster (Figure 4.25). Using this strain with the same xylose concentration used for the split cluster strain resulted in no significant growth rate differences compared to the wild type in any of the conditions we tested (Figure 4.27). We observed the mutant in the amino acid supplemented and deprived media but saw no obvious morphological defects that might explain the slow growth phenotype.



56 - Figure 4.27 - Effect of disrupted gene co-localization in different media
Investigating the role of nutrients in a TCE disrupted mutant. Error bars are the standard deviation of each sample. All MOPS media contain tryptophan as the *B. subtilis* strain is auxotroph and glucose as the carbon source. The 6 amino supplementation includes methionine, histidine, arginine, proline, threonine and tryptophan.

Splitting the cluster causes genome wide expression changes

To better understand the role of arginine in the context of the cluster we measured the transcript levels of several genes of the cluster using qPCR. We found that in the wild type, there was very little difference in the expression of the cluster genes between the two media (Figure 4.28). However, even though only the latter four genes were genetically manipulated we saw an increase in expression of *pyrH* and *frr* in the split cluster mutant (Figure 4.28). This suggests that either the antibiotic cassette placed downstream of these genes is influencing the expression or that there is a feedback mechanism responding to the changes caused by the translocation of the cluster. As the antibiotic cassette is placed in the reverse orientation in relation to the other genes and is followed by several terminators to insulate its expression it is unlikely this would be the case of an increase in expression. In the arginine depleted media, we see a striking decrease (9-fold) in *uppS* expression (Figure 4.28) even though induction of xylose remained the same in this media condition. Furthermore, *proS,* which was under the control of the same promoter, did not share the same behaviour. We saw previously that the *uppS* UTR did not show reduced activity in the same media conditions in our promoter fusion experiments conflicting with this observation (Figure 4.11). At this stage we do not have an explanation for the drop in *uppS* levels in this condition.

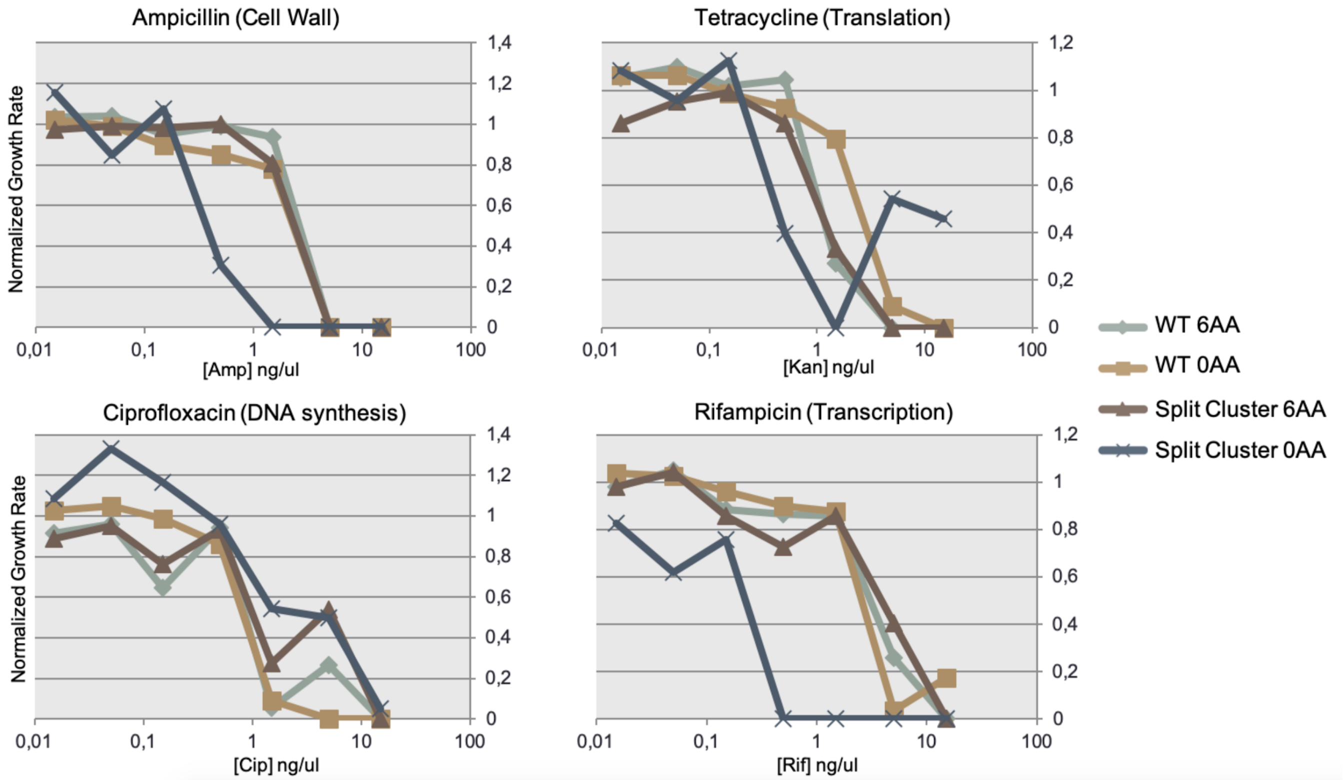


57 - Figure 4.28 - qPCR of the cluster genes in the different conditions and strains

Transcript abundance of the TCE cluster genes in different media and genetic contexts measured with qPCR. The values are fold changes in relation to the wildtype in MOPS media with no amino acid supplementation. The error bars represent the standard error of the replicates. qPCR levels were normalized to the constitutively expressed genes *recA* and *gyrB*.

Antibiotic sensitivity of the split cluster

In order to observe how else, the disruption of the eight gene-cluster may have perturbed other biological processes in the cell we applied four antibiotics: ampicillin, tetracycline, ciprofloxacin and rifampicin to target cell wall synthesis, translation, DNA synthesis and transcription respectively. In amino acid supplemented conditions, we observed no difference in antibiotic sensitivity between the mutant and wild type (Figure 4.29). It was seen, however, that in the amino acid deprived conditions there was a 10-fold increase in sensitivity to both ampicillin and rifampicin of the mutant strain (Figure 4.29). That the cell is more sensitive to cell wall perturbing antibiotics is not unexpected as *uppS* is under the control of a xylose inducible promoter which was at the same level in both the amino acid present and absent conditions, and thereby incorrect tuning by this promoter might be the cause of the increased sensitivity. From the qPCR experiments we saw that *uppS* in the 0 amino acid condition may be repressed resulting in less peptidoglycan synthesis and a susceptible cell wall (Figure 4.28). Why the cell may be more sensitive to transcription inhibition is an open question. We have speculated that splitting the cluster could disrupt pyrimidine biosynthesis which may inhibit UTP production and therefore its availability to RNA polymerase during transcription. One possibility, however, is that as only the mutant in the amino acid absent condition grows at a slower rate, potentially it is the slower growth that results in increased sensitivity and not the disruption of the cluster directly.



58 - Figure 4.29 - Effect of disrupted TCE co-localization on antibiotic sensitivity
Here we observe antibiotic sensitivity of *B.subtilis* in different media in the context of TCE cluster disruption. Strains were grown in MOPS + glucose and antibiotics were added during exponential phase. Growth rates are normalized to the growth rate of the strain in the media condition with no antibiotic addition. Growth rates were taken 2 hours after addition of the antibiotics.

Mutant recovers auxotrophic phenotype via the xylose inducible promoters

Throughout our handling of the split cluster mutant strain we observed that long incubations in minimal media (without amino acid supplementation) resulted in eventual death of the cell. However, in some cases we noted that regrowth occurred after long incubation time. Further inoculation of these cultures revealed a recovery to the wild type growth rate in the minimal media. We postulated these samples may have undergone a compensatory mutation and therefore submitted a sample for DNA sequencing alongside our wild type strain. We discovered there were several intergenic SNPs (n = 38) within our lab strain compared to the published W168 genome and in the mutant strain we found an additional 8 SNPs (Table 4.4). Relevant to the phenotypes we observed, we saw mutations in the *frr* and *xylR* genes. The mutations in the frr gene stems from our cloning method which relies on Golden Gate cloning using the type II restriction enzymes *Bsa*I and *Bpi*I. In order to create constructs, we needed to ensure there were no restriction sites for these two enzymes, and in the homology region which covered the frr region there was one. We removed it by making a synonymous nucleotide substitution. Both the translocated *uppS* and the downstream *dxr* gene are regulated by the xylose inducible promoter in the mutant. The SNP in the *xylR* gene results in a frameshift which most likely completely destroys the function of the encoded protein. The xylose inducible promoter P_{xylA} is controlled by upstream operator sites (Pxylo) which in the absence of xylose are normally bound by XylR repression expression of the promoter (Gärtner et al. 1992). In the presence of xylose XylR dissociates from the operator site allowing expression. A non-functional *xylR* gene would result in no repression of the xylose promoters. As these mutations occur after overnight cultures where the supplied xylose has either been metabolized or diluted between many cells perhaps these mutations arose due low xylose levels resulting in low *uppS* and *proS* expression, which we know results in low growth rates or lethality, and not from splitting the cluster itself. We also saw two missense SNPs in *cypA* and SNPs in the ribosomal RNA genes *rrnB* and *rrnI*. *cypA* is a cytochrome P450-like enzyme which is involved in the detoxification and degradation of polychlorinated biphenyls (Sun, Pan, and Zhu 2018). It is unlikely these mutations are relevant to the phenotype we observed since PCBs are not naturally occurring compounds and would not be present in the media.

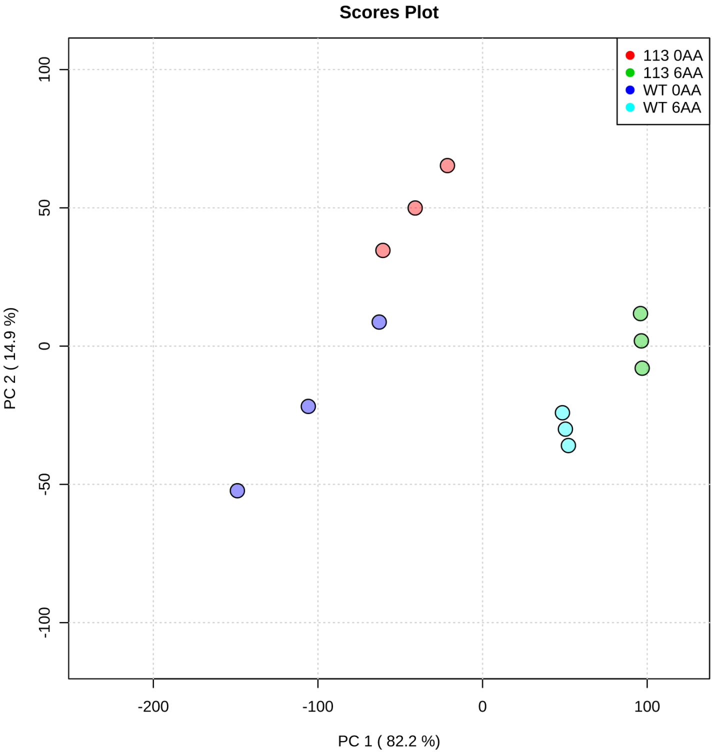
Gene Name	Frame Shift	Missense	Stop Gained	Synonymous
cypA	0	2	0	1
frr	0	0	0	2
rrnB-16S	0	1	0	0
rrnI-16S	0	1	0	0
xylR	1	0	0	0
rluD	0	1/2*	0	0
rrnH-23S	0	7	0	0
rrnG-16S	0	2	0	0
comP	0	1	0	0
epsC	0	1	0	0
gerAA	0	1	0	0
gltA	0	1	0	0
ilvC	0	0	0	1
narG	0	1	0	0
oppD	0	1	0	0
pgdS	0	0	0	1
pksN	1	0	0	0
prpC	0	1	0	0
rluB	0	0	0	1
rplW	0	1	0	0
sacA	0	1	0	0
sepF	0	1	0	0
sigI	0	1	0	0
trmD	0	1	0	0
yesY	1	0	0	0
yheH	0	0	1	0
yjcM	0	1	0	0
ymfD	0	1	0	0
yoqA	0	1	0	0
yozT	0	0	0	1
yqeZ	0	0	0	1
ytpS	0	1	0	0
yulF	0	1	0	0
yutE	0	1	0	0
yxbD	0	1	0	0

6 - Table 4.4 - Genome sequencing of the mutant

The split cluster mutant was allowed to grow overnight in the MOPS media without amino acid supplementation. The culture was grown to confirm the recovery of growth rate and the culture was streaked and DNA sequenced. The bolded and underlined genes are rows where the SNPs were found only in the mutant strain, all other SNPs were found in both samples. *_rluD_ was the only exception and was found to have a SNP less in the mutant strain, the first number represents SNPs discovered in wild type and the second, the number in the mutants. rRNA genes were classified as missense mutations despite not encoding an amino acid sequence.

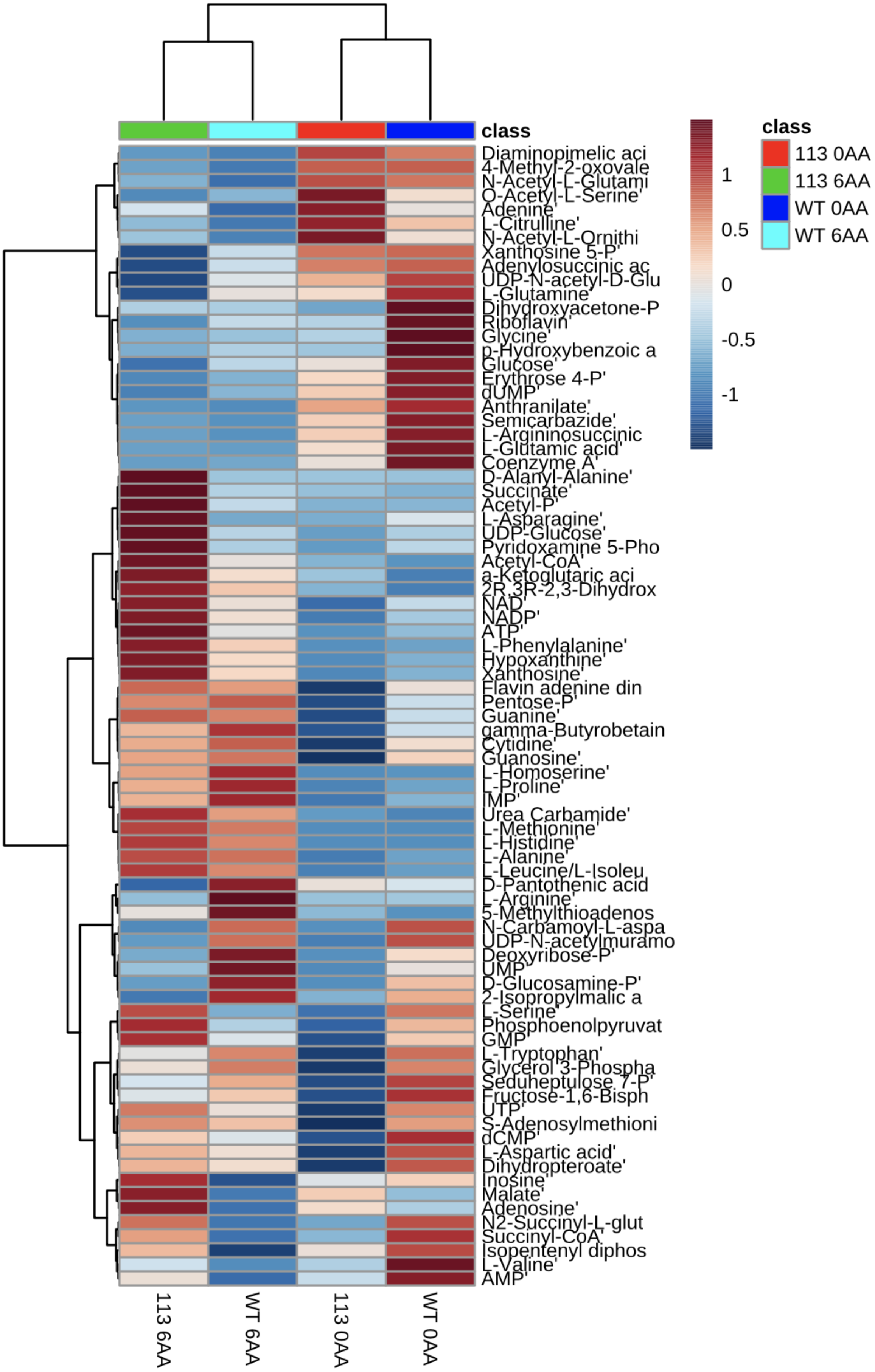
4.10 Metabolomic profile of the TCE split cluster strain

Our analysis of the TCE cluster has pointed to several possible perturbations of key metabolic pathways including arginine and pyrimidine biosynthesis and central carbon metabolism. Therefore, we performed metabolite extractions of both the wild type and split cluster mutant in MOPS media, amino acid lacking and supplemented. Our collaborators in the Link group (Stefano Donati) performed targeted mass spectrometry to acquire relative intracellular concentrations of several metabolites between our samples. We performed a principal component analysis of the 81 measured metabolites. The analysis revealed that the samples clustered together based on the media the sample was collected from more than the genetic context of the organism with most of the variance being explained in PC1 (82.2%) (Figure 4.30). This is unsurprising as in the absence of amino acids the bacteria would need to *de novo* synthesize them, resulting in large shifts in the metabolism and metabolites. Variability between replicates was higher in the 0 amino acid condition (Figure 4.30).



59 - Figure 4.30 - Metabolite differences caused by media conditions is greater than genetic differences

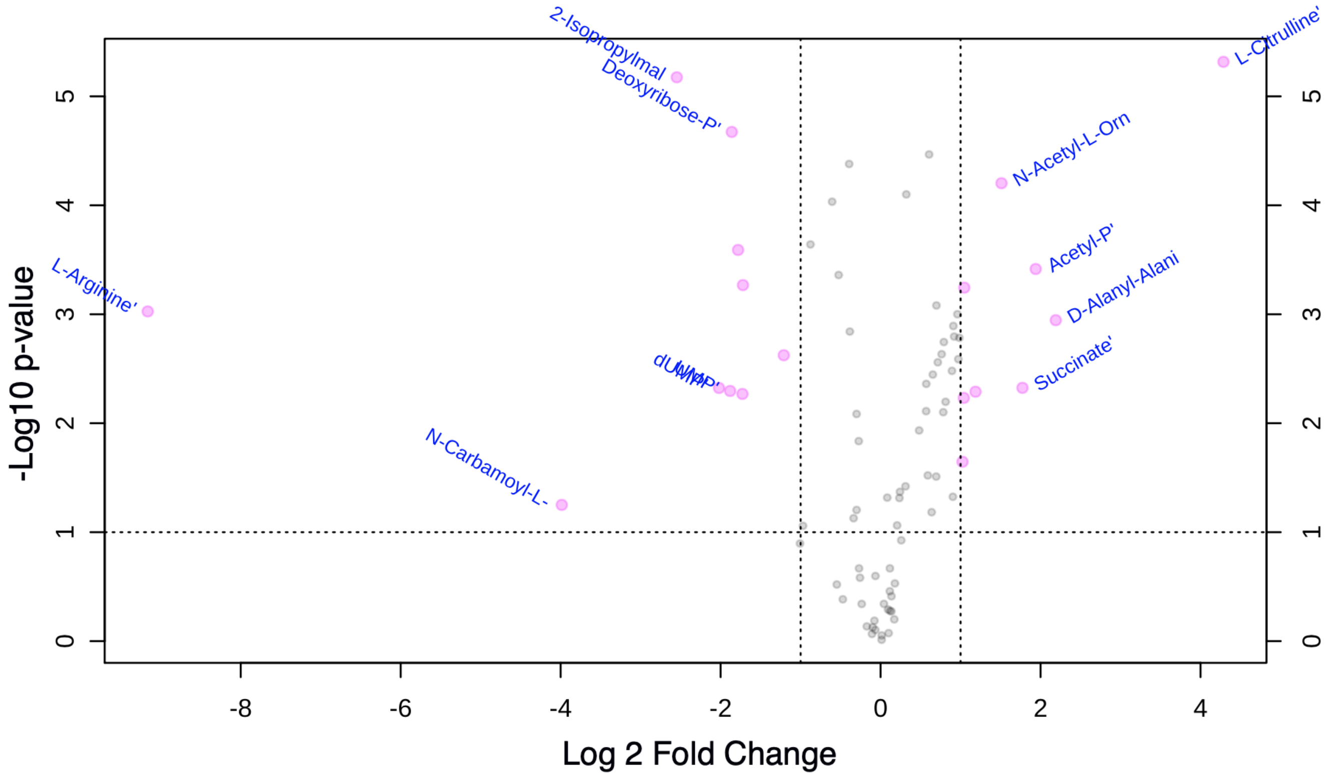
Principal component analysis of four conditions which compared metabolite concentration between the wild type of Bacillus subtilis and a mutant where the TCE cluster was de-localized in both minimal and amino acid supplemented media. Amino acid supplementation consisted of tryptophan, arginine, proline, serine, threonine and methionine. Each point represents a different sample and the colour represents the strain and condition. The point in the middle is an average of the two samples per condition.


class
Diaminopimelic aci
4-Methyl-2-oxovale
N-Acetyl-L-Glutami
O-Acetyl-L-Serine'
Adenine'
L-Citrulline'
N-Acetyl-L-Ornithi
Xanthosine 5-P'
Adenylosuccinic ac
UDP-N-acetyl-D-Glu
L-Glutamine'
Dihydroxyacetone-P
Riboflavin'
Glycine'
p-Hydroxybenzoic a
Glucose'
Erythrose 4-P'
dUMP'
Anthranilate'
Semicarbazide'
L-Argininosuccinic
L-Glutamic acid'
Coenzyme A'
D-Alanyl-Alanine'
Succinate'
Acetyl-P'
L-Asparagine'
UDP-Glucose'
Pyridoxamine 5-Pho
Acetyl-CoA'
a-Ketoglutaric aci
2R,3R-2,3-Dihydrox
NAD'
NADP'
ATP'
L-Phenylalanine'
Hypoxanthine'
Xanthosine'
Flavin adenine din
Pentose-P'
Guanine'
gamma-Butyrobetain
Cytidine'
Guanosine'
L-Homoserine'
L-Proline'
IMP'
Urea Carbamide'
L-Methionine'
L-Histidine'
L-Alanine'
L-Leucine/L-Isoleu
D-Pantothenic acid
L-Arginine'
5-Methylthioadenos
N-Carbamoyl-L-aspa
UDP-N-acetylmuramo
Deoxyribose-P'
UMP'
D-Glucosamine-P'
2-Isopropylmalic a
L-Serine'
Phosphoenolpyruvat
GMP'
L-Tryptophan'
Glycerol 3-Phospha
Seduheptulose 7-P'
Fructose-1,6-Bisph
UTP'
S-Adenosylmethioni
dCMP'
L-Aspartic acid'
Dihydropteroate'
Inosine'
Malate'
Adenosine'
N2-Succinyl-L-glut
Succinyl-CoA'
Isopentenyl diphos
L-Valine'
AMP'
113 6AA
WT 6AA
113 0AA
WT 0AA
1
0.5
0
-0.5
-1
class
113 0AA
113 6AA
WT 0AA
WT 6AA

60 - Figure 4.31 - Metabolic changes between the conditions

A heatmap showing the relative abundances of the different targeted metabolites in the four different conditions. The colour is an average of the different replicates for each condition. 113 is the ID number for the split cluster strain.

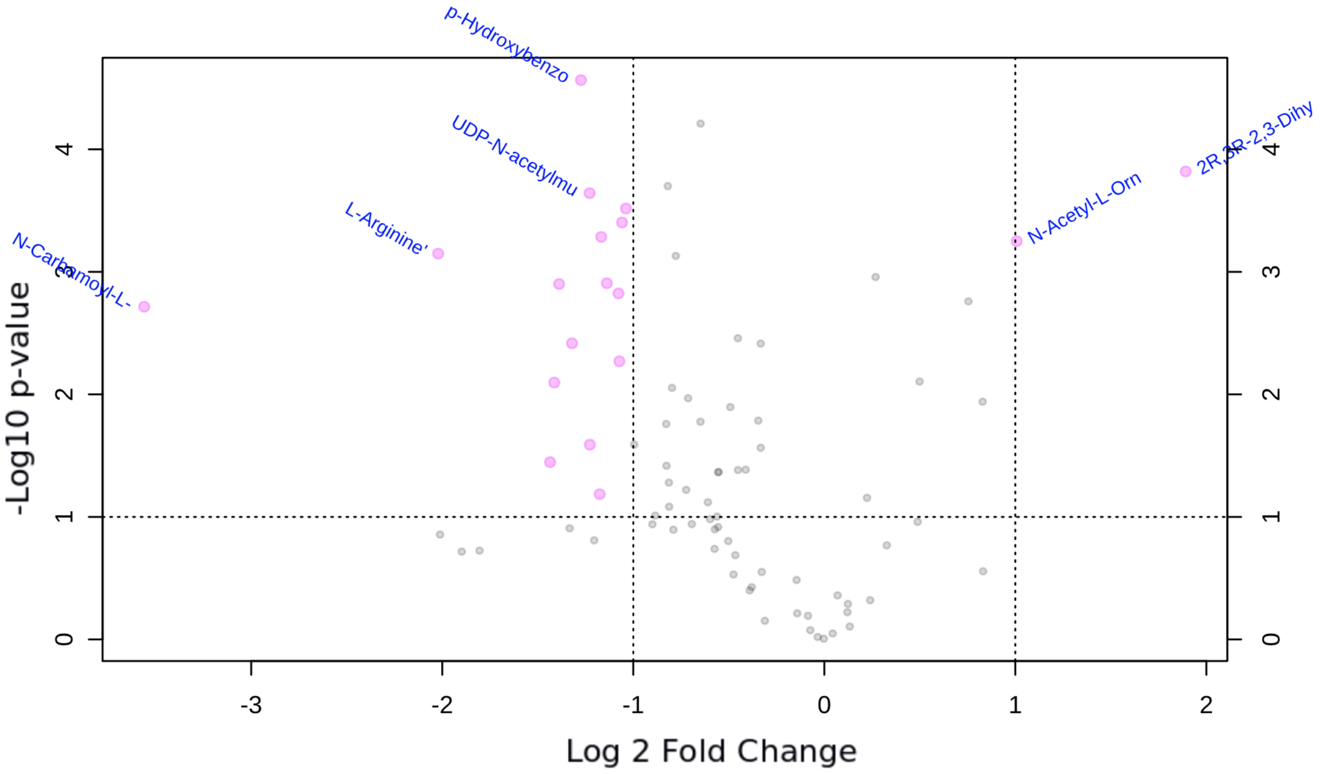
When comparing the split cluster strain to the wild type in the presence of amino acids we saw 10 metabolites that were significantly lower in abundance in the mutant (Figure 4.32). Notably several of these belonged to the aforementioned pyrimidine and arginine metabolic pathways including arginine, N-Carbamoyl-L-aspartate', dUMP, UMP and UDP-N-acetyl-D-Glucosamine. Conversely some metabolites in this pathway had increased concentrations including N-Acetyl-L-Ornithine, L-Citrulline and N-Acetyl-L-Glutamic acid. We also found many metabolites related to the TCA cycle to have significant differences in pool levels including: 2-Isopropylmalic acid, a derivative from the TCA intermediate succinic acid; D-Glucosamine-P which is synthesized from glutamine and pantothenic acid which is a precursor to the synthesis of coenzyme A which were all less present. The TCA intermediates succinitate, acetyl-CoA and asparagine, a derivative for oxaloacetate, were higher in concentration. We also observed that the split cluster had higher concentrations of ATP than the wild type with amino acid supplementation.



61 - Figure 4.32 - Metabolite abundance differences with amino acid supplementation

A volcano plot comparing the mutant and the wild type in MOPS minimal media supplemented with 6 amino acids. Points in pink represent metabolites that were over the significance and fold change thresholds. Fold change is mutant/wt, therefore metabolites which are present in higher levels in the wild type appear on the left side of the figure.

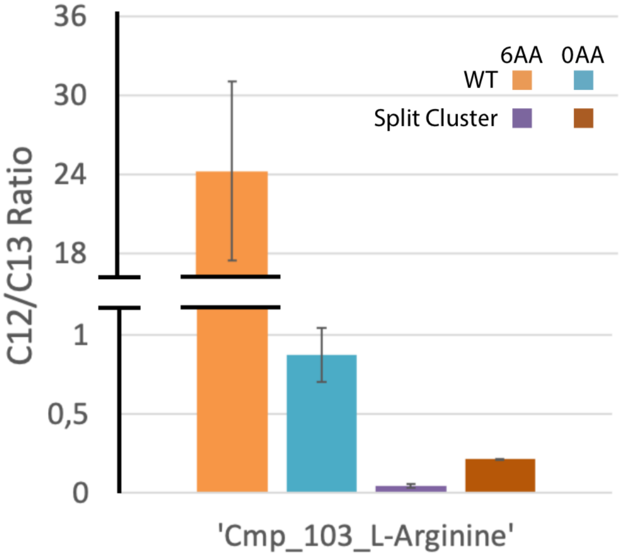
In the amino acid lacking conditions we saw that the mutant had very few metabolites in higher abundance than the wild type with only two showing a significant increase, ornithine a precursor and catabolic product of arginine, and 2R,3R-2,3-Dihydroxy-3-methylpentanoate which is involved in valine, leucine and isoleucine biosynthesis (Figure 4.33). As in the amino acid supplemented media, we saw decreases in the mutant for carbamoyl-aspartate, arginine and glucosamine. Several different and diverse metabolites were lower in the mutant in this condition including cytidine, dihydropteroate, NAD, guanosine, UDP-N-acetylmuramoyl-L-alanine and Deoxyribose-P (down in both conditions) (Figure 4.33).



62 - Figure 4.33 - Metabolite abundance differences without amino acid supplementation

A volcano plot comparing the mutant and the wild type in MOPS minimal media not supplemented with amino acids. Points in pink represent metabolites that were over the significance and fold change thresholds. Fold change is mutant/wt, therefore metabolites which are present in higher levels in the wild type appear on the left side of the figure.

A notable metabolite to observe is arginine. Arginine is one of the amino acids we supply in the 6 amino acid containing media. *B. subtilis* imports available arginine as can be seen in wild type by its high intracellular concentrations (Figure 4.34). When provided with no arginine in the media, we saw the wild type strain had significantly lower concentrations of arginine. As the wild type cannot import arginine in this condition it must rely on *de novo* biosynthesis. We also see low levels of arginine in the split cluster mutant when there is no arginine in the media, however, unlike the wild type when arginine is provided to the split cluster there is no increase in arginine intracellular concentrations (Figure 4.34). Therefore, we could postulate two reasons to explain this behaviour: either arginine import has been blocked through an unknown mechanism in the mutant leaving the cell to rely on *de novo* biosynthesis for its arginine requirements or that arginine consumption is drastically increased. If arginine is the only limiting factor in the split cluster, as evidenced by the fact we can restore the normal growth phenotype with the addition of arginine, the first possibility is unlikely unless *de novo* biosynthesis of arginine is also in some way hindered. The other three of the other amino acids that were supplied, methionine, proline and histidine, behaved as we would expect.

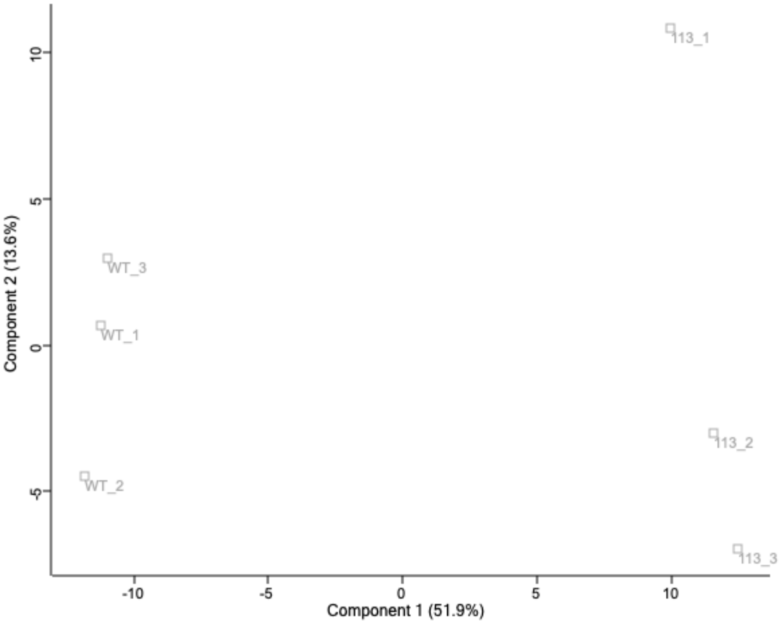


63 - Figure 4.34 Arginine was not present in the split cluster strain

Targeted mass spectrometry C12/C13 ratios for arginine. Error bars represent the standard deviation between the replicates. The y-axis has been split to show the large increase in arginine in the wild type whilst still highlighting the differences in the other samples.

4.11 Proteome of the TCE split cluster

We performed protein extractions of the wild type and split cluster mutant from exponentially growing cultures in the amino acid supplemented MOPS media and were able to get measurements for 1318 proteins. We were able to see clear clustering between the samples on the first component of a principal component analysis (Figure 4.35). Between the samples we found many proteins with significant differences in expression (n = 307) with the majority being lower in the mutant (n = 219) rather than higher (n = 89) (Figure 4.36 and Table 8.10). Of the eight cluster genes we were only unable to get measurements for *cdsA*, most likely a result of our implemented protein extraction strategy which is known to have issues with membrane proteins. Two of the genes were also shown to have significant differences, *tsf* and *uppS*. *tsf* is higher in the wild type with a difference of 0.348 whereas *uppS* is higher in the mutant with a difference of -1.252. This does counter our qPCR findings in the mutant strain in the same media conditions where we saw lower levels of *uppS* transcripts compared to the wild type (Figure 4.28). As the standard deviation between the replicates in the qPCR experiment was very high, we could have overestimated the reduction in *uppS* levels. Since we controlled expression of UppS expression through a xylose inducible promoter and optimized the xylose concentration for optimum growth rates in the media, a higher level of UppS could suggest that it may be less effective in its new context as it was needed to have a higher expression to achieve the same growth rate.



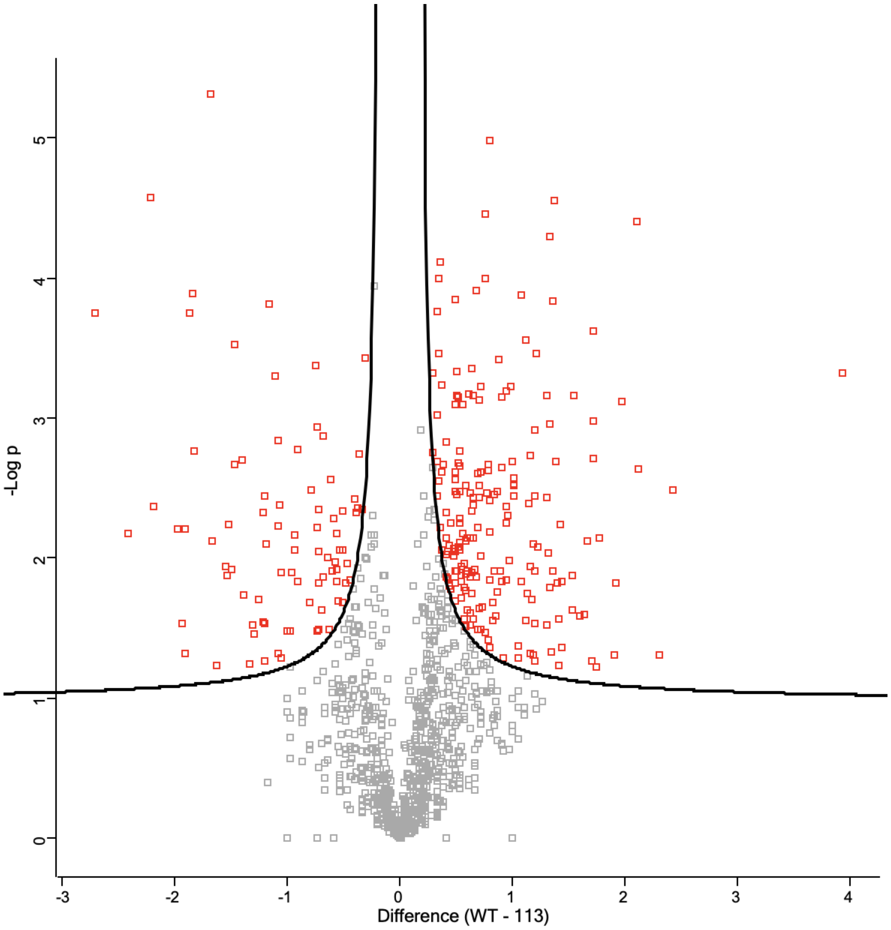
64 - Figure 4.35 - Protein samples are similar based on strain

A PCA plot of the metabolite data. Each point represents a different sample. Proportion of variance for the two components are written as percentages next to the axis labels. Samples were grown in MOPS media supplemented with amino acids.

To improve our understanding on what these major shifts in protein levels could mean, we subjected the proteins with significant differenences between the two samples to enrichment analysis, using all detected proteins as the background. Processes we saw upregulated in the mutant included many ribosomal proteins, proteins involved in the TCA cycle, membrane proteins and those involved signal transduction. Given there is a major shift in protein expression between these two samples it is unsurprising we would see an enrichment in signal conveying proteins. Among them are many transcriptional factors, notably the purine operon repressor (PurR), the regulator of genetic competence and quorum sensing (ComA) and the transcriptional regulator of transition state (AbrB). We also saw the upregulation of the transcriptional regulator, PsdR (formally YvcP) which is induced by lipid II-binding lantibiotics, such as nisin which is known to subsequently induce the Psd system to export the toxic peptides (Zhang et al. 2016). As in the metabolomic data, we saw several proteins from the TCA were upregulated including FumC, PdhC and SucD responsible for the synthesis of malate, acetyl-CoA and succinate respectively (Figure 4.36) with the latter two also showing an increase of the corresponding metabolite and the former not being measured (Figure 4.32). Other proteins involved in carbon metabolism that were upregulated in the mutant include DapB, IlvH, LysC, DhbC, GlmM, Prs and Tkt.

Amongst the proteins less expressed in the mutant we found enrichments in pyrimidine biosynthesis, diaminopimelate biosynthesis, the TCA cycle, and mobility (Figure 4.36) as well as the sigma factors σ^A and σ^B. Within the TCA cycle both citrate synthases and isocitrate dehydrogenase were downregulated. Citrate was not measured in our metabolomic experiments (Figure 4.32), however the product of isocitrate dehydrogenase, ɑ-ketoglutarate, was found to be higher in concentration, suggesting this metabolite may be building up. Regarding pyrimidine biosynthesis, dihydroorotase (PyrC), dihydroorotate dehydrogenase B (PyrK), orotidine 5'-phosphate decarboxylase (PyrF) were downregulated. This corresponds with the metabolic experiments as PyrF catalyzes the decarboxylation of orotidine 5'-monophosphate (OMP) to uridine 5'-monophosphate (UMP) and we see the decrease in enzyme concentration correlates with decrease in UMP concentrations in the metabolomic data. This was also the case for PyrC which catalyzes the reversible cyclization of carbamoyl aspartate to dihydroorotate. We did not see similar decreases in the protein levels for the rest of the operon which is also controlled by the same σ^A

dependant promoter. PyrR is found at the start of the operon and has RNA binding activity which is stimulated by UMP and UTP and results in the transcription termination by the binding of anti-antiterminators (Hobl and Mack 2007). We also saw a downregulation of the σ^A regulated glutamine synthetase (GlnA), although we did not get measurements for its two negative regulators GlnR and TnrA. Reduced levels of GlnA would result in higher glutamate pool levels and therefore also a buildup of α-ketoglutarate, which could explain the build-up of this metabolite despite the downregulation of isocitrate dehydrogenase. We also see the significant downregulation of YhdL. YhdL is an anti-σ^M factor, implying the cell may be compensating for cell wall stress and that $\sigma^M$is more active in the cell.



65 - Figure 4.36 - Differentially expressed proteins between the wild type and split cluster mutant

A volcano plot showing the proteins which were differentially expressed between the wild type and split cluster mutant in MOPS with amino acid supplementation. The x axis represents the difference in $\log_2$ signal strength (thereby $\log_2$ fold change). Proteins which are more highly expressed in the mutant are on the left of the plot. Red points indicate proteins which were considered significantly differentially expressed, passing the significance and difference thresholds which are represented by the black line.

RelA, an important protein in stringent response, was shown to be less prevalent in the split cluster strain. Mutants lacking RelA are still viable but are unable to undergo the stringent response (Fiil and Friesen 1968). The stringent response results in a downregulation of resource-consuming cell processes such as transcription and translation whilst simultaneously upregulating the expression of biosynthetic genes (Jain, Kumar, and Chatterji 2006; Chatterji and Ojha 2001). This is in agreement with our data as most upregulated proteins were associated with ribosomes or translation. A lack of amino acids is known to trigger the starvation response in bacteria when the ribosome encounters deacylated tRNA in the ribosomal A-site. To see if the split cluster is more sensitive to the stringent response, we supplied serine hydroxamate (SHX) to the media at varying concentrations. SHX is a serine mimic that triggers the stringent response (Tosa and Pizer 1971). As stringent response greatly limits the growth rate of the cell, this could be an explanation for why we see a decrease in growth rate of the split cluster when no amino acids are present. Therefore, we would expect the cell to be more sensitive to SHX if the stringent response is being triggered early in media lacking amino acids in the mutant. In preliminary experiments we saw that in amino acid supplemented media both the WT and split cluster strains were affected in the same way to different SHX concentrations, implying that their stringent response and recognition is the same. This also suggests the slow growth is not due to the stringent response being activated.

4.12 Summary

In this chapter, we characterized the Translation-Cell Envelope gene cluster which we identified using GenCoDB as an interesting candidate for synchronizing cell volume and cell envelope synthesis. The cluster consists of many essential genes holding key rate limiting steps in their respective pathways and was found to be broadly conserved across the bacterial kingdom, suggesting that it may fulfil a role as a synchronizer between the many processes contained in the cluster. We discerned that polycistronic transcripts comprising of the translation and cell envelope fractions of the cluster are inconsequential *in vivo*. This was despite previous experimental evidence, tight transcriptional correlation between the genes, and suggestions from the literature. Through other experiments we were able to confirm that co-localization of *rasP* in the TCE cluster was not relevant for cell size homeostasis nor is it likely to contribute to the localization of the genome to the membrane through transertion.

We investigated the impact pyrimidine and arginine metabolism on the cell by analysing a mutant where the cluster was split in half. This revealed a dependency of arginine-like amino acids which appeared to be rapidly catabolized in the cell resulting in growth deficiency in arginine lacking media. Examining the changes in the metabolome and proteome of these mutants confirmed the impact the delocalization of these genes had on both pyrimidine and arginine *de novo* biosynthesis and highlighted the subsequent effects on central carbon metabolism.

5. Discussions and Conclusions

Bacteria are among the simplest organisms on our planet, and yet the question of how they orchestrate their individual processes to maintain stable growth is still a question out of our reach. One specific question is how the cell regulates the increase in volume and surface area that comes with increased growth rates. As introduced in Chapter 1, a new avenue to study this topic is from the context of conserved gene neighbourhoods. Doing so requires a platform allowing both the quantitative and statistical analysis of gene neighbourhoods and an understanding of their behaviour in the context of genome evolution. In this chapter we will discuss the possible: interpretations, implications, limitations and future recommendations for the data we presented in the last three chapters. First, we will discuss GenCoDB and how our choices in species bias correction and significance adds value to the microbiology community in comparison with other strategies. Next we will discuss our analysis on the evolution of gene clusters in the context of the selective pressures that maintain them. Finally, we will examine our findings of a physiological benefit of a non-canonical gene cluster which brings together ribosome-associated and cell envelope genes.

5.1 GenCoDB

Our goal with the development of GenCoDB was to establish a tool solely for the purpose of finding gene neighbourhoods linking ribosome-associated genes with cell envelope biosynthesis. In the process of creating this tool, we ended up creating one with much broader applicability which distinguishes itself amongst other bacterial genome comparison web tools. First and foremost, GenCoDB can be used to analyse conserved gene neighbourhoods, facilitating future research in this area. It is of particular use for laboratories lacking bioinformatic support or that currently rely on non-quantitative tools such as MicrobesOnline. One area positively impacted by the access of GenCoDB is research into extra cytoplasmic function (ECF) sigma factors which are a cellular mechanism for signal transduction in bacteria often found co-localized with their direct target genes, genes encoding functions important for the signal transduction mechanism or their anti-σ factors (Jordan et al. 2006; Joseph et al. 2002). Both Staroń et al. and X. Huang et al. used genome context data to characterize their ECF classification groups using MicrobesOnline (Dehal et al. 2010) with the former using additionally using THE SEED (Ross Overbeek et al. 2005) - a web tool that is no longer available. Using MicrobesOnline they could only count the raw frequency of genes appeared in the genomic context with no attention towards the significance or bias their genome subsets could be providing. More recently in another field, Szadkowski et al. were looking for additional components of a protein module of interest (the MglA-MglB-RomR module). The authors used a mixture of BlastP and custom Perl scripts in order to find that the protein RomX co-localized with this module. They then experimentally verified the importance of the co-localization of RomX in the lab, as a critical interaction partner of RomR. In all of these cases, if they had had access to GenCoDB this would have provided them with quicker, and more quantitative results.

In Chapter 2, as part of identifying significantly conserved ortholog groups, we measured the average neighbourhood conservation for many different taxonomic subsets (Figure 2.3). Generally, the conservation correlated with the genetic diversity within the group. When genetic diversity was low there was also an increase in the variance of average neighbourhood conservation values. This can be understood due to the stochasticity inherent in genomic rearrangement events and the heightened effect of noise due to the smaller sample sizes of these clades. There were a few notable exceptions where median conservation was higher or lower than the model. Those which had higher levels of conservation relative to their genetic diversity included Spirochaeta, Mollicutes, and Cytophagales. Both Mollicutes are known to have highly reduced genomes which would reduce the number of possible genome permutations independent of rRNA sequence differences explaining their trend towards more conserved neighbourhoods (Sirand-Pugnet et al. 2007). Furthermore, mycoplasmataceae (a family within Mollicutes) species specifically have been shown to have reduced evidence of inversions occurring on their genome, drastically minimizing the effect of chromosome shuffling on gene order (Suyama and Bork 2001). Previous work on gene cluster analysis in Mollicutes reported they had the smallest number of gene clusters when compared to other bacteria; however, this did not factor in the smaller genome size nor the genetic diversity of the Mollicutes to the other bacterial subgroups (Y. Zheng et al. 2005). As to the other clades we could not find literature evidence suggesting why they have higher levels of neighbourhood conservation; however, there are many reasons that could lead to this observation. For instance, a lack of restriction enzymes that could result in strand breaks in the DNA or increasing the likelihood of errors during repair. The actinobacteria phylum was shown to have less conserved neighbourhoods relative to their diversity. Counterintuitively, actinobacteria engage less with genome reorganizing HGT events than any other phylum (Lewin et al. 2016), however, the average genome sizes are significantly larger (on average greater than 5 megabytes) (Barka et al. 2016). Larger genomes correspond to larger gene repertoires and signify that several gene duplication events must have occurred if the acquisition of genes was not through HGT resulting in a shuffled genome.

With the development of our method, we were aware of two challenges, how to overcome the genome sampling bias and determining significance in conservation. As we established earlier, the significance that two genes appear co-localized is dependent on the context of which genomes are being analysed. Namely co-localization seen in species closely related to each other is less significant because there is less evolutionary time between them in which rearrangement events could occur. We approached this problem with a methodical approach that measured the expected neighbourhood conservation at different subsets of genomes (taxonomic divisions) and used this as a basis of defining a threshold of significance (Figure 2.3). This relationship fit tightly at large scale divisions with only a few minor outliers that could be explained due to larger and smaller average genome sizes of the clades. Our model only began to break down in smaller taxonomic groups with higher noise and smaller sample sizes which hamper analysis (Figure 2.3). Depending on the resources used to study neighbourhood conservation, the way this challenge is approached varies. In genome browser based comparison tools (e.g MicrobesOnline (Dehal et al. 2010) and the JGI genome portal (Grigoriev et al. 2012) where the neighbourhoods are aligned centred on one ortholog group, quantitative statistics are not provided; subsequently, it is impossible to determine if an observation is significant. Therefore, users would need to perform their own downstream analysis. StringDB does not directly provide significance to their genomic context information and instead provides a combined score which is the combination of many factors, co-localization included (Snel et al. 2000). As well as co-localization, these factors include text mining, gene fusion, co-expression co-occurrence. For co-localization they only consider genes with intergenic distances of less than 300 nucleotides (Snel et al. 2000). To generate the scores the number of occurrences of a gene pair is compared to the chance that this gene pair would appear together in randomly shuffled genomes (Snel et al. 2000). This probability of a gene pair occurring in two genomes was calculated 0.02, decreases to <0.002 for three species and <0.0005 for four species. As we discussed in Section 2.2, basing co-localization statistics of randomized genomes is a vast underrepresentation of the true expected co-localization frequency in real bacterial datasets. This is because bacterial genomes are not independent from each other and share significant similarities especially at smaller evolutionary distances. In the clustering work performed by R. Overbeek et al., they implemented an arbitrary threshold to the sum of "pair of close bidirectional best hits" scores they assigned to each of their gene pairs. Overlapping pairs with scores higher than this threshold would iteratively be clustered together to classify clusters. The threshold was then optimized to get a clear disentanglement of clusters. Fang, Rocha, and Danchin, implemented a more sophisticated statistical test in determining the significance of gene pairs, the Kuiper test. This tests against the null hypothesis that a given distribution is uniformly distributed, and in this case the distribution consisted of the gene distances between a gene pair in the surveyed genomes. This test is independent of the gene distance, therefore would not exclude if a gene pair is constantly half a genome apart from one another. However, we would consider this scenario extremely uncommon. Aside from the minor filtering they do to remove extremely distant and closely related species (explained in the next section) the distance between genes that are from a sister clade and those from a different phylum are treated identical and thereby the statistical weight that comes with such an observation is not included. None of these methods aside from that implemented in R. Overbeek et al. 1999 takes the genetic diversity strongly into account when modulating how they define significance and therefore, with fine tuning of the selected genome pool, any gene pair could be arbitrarily made to be significant. R. Overbeek et al. calculated the score of a gene pair being directly proportional to the genome distance. Therefore, a genome subset consisting of closely related genomes would surpass the threshold. However, as this threshold is arbitrarily chosen and static it would not be able to easily adapt to a changing genome dataset. We therefore believe our method of significance is a good solution to the rapidly changing and expanding the availability of bacterial genomes.

Despite the efforts of movements such as the Tree of Life, which aim for diversity in genome sequencing, our need to understand pathogenic and economically relevant species has resulted in us favouring acquiring genomes non-uniformly across the bacterial kingdom. In extreme cases some bacterial datasets have several sequences of one species, usually in model organisms such as *E. coli* or *B. subtilis (Chen et al. 2019; Kriventseva et al. 2015; Dehal et al. 2010)*. In the case of genome comparisons, the presence of these extra genomes, or over/under-sampling of certain claims results in a bias and ultimately in false positives and negatives. We reduce the impact of this bias with an elegant method that scales the contribution of each genome relative to its similarity to other species in the dataset based on a 16S rRNA tree (Figure 2.2). This method was similar to what was performed in Vieira-Silva and Rocha, 2010, and does have limitations which we will outline later in this chapter. This method is advantageous as it adapts to the genomes it has available and therefore allows the database to upscale without the risk of sampling bias changing the interpretation. To the best of our knowledge, GenCoDB is the only bacterial genome comparative website which takes this bias into account. Many databases, especially those which do not provide quantitative statistics generally include every genome that meets their quality thresholds (Dehal et al. 2010; Chen et al. 2019). Other databases perform curation steps before including their genomes with the aim of including only representative species, thereby minimizing the effect of genomes from multiple strains of an organism (Kriventseva et al. 2019; Szklarczyk et al. 2019). Crucially this curation was performed in the most recent version of OrthoDB from which we decided which genomes to include in our dataset. We would argue that this effort is insufficient. Firstly, it is manual and time consuming, thereby not allowing it to be scalable. Secondly, without any objective measures to define genetic distance, it is impossible to know exactly how over-represented a clade may be. We saw this in our OrthoDB dataset, where although proteobacteria made up a significant proportion of the genome sequences and actinobacteria less so, analysis of the genetic diversity of these phyla showed that actinobacteria are not very diverse and the contribution should decrease and vice versa for the proteobacteria (Figure 2.2). Other work which identified clusters in bacterial genomes utilized different methods in order to overcome these challenges. As already explained R. Overbeek et al. uses a similar method as the one we implemented by utilizing the phylogenetic distance on a 16S rRNA tree to score their gene pairs. To minimize the influence generated by many closely related species (such as strains) and from extremely distant species, Fang, Rocha, and Danchin excluded the top and bottom 10% gene distances scores they generated for each gene pair. This is similar to our method at attributing contribution values to genomes, where closely related species are assigned contribution values that sum to a total of 1, and extremely distant species can only contribute up to a maximum of 1. It is different, however, in that the method implemented in Fang, Rocha, and Danchin, 2008 relies on the assumption that all the similar (or distant) sequences are adequately captured in the 10% window, whereas in reality the window could be much larger or smaller. In case it is smaller, this would be introducing a bias by removing perfectly valid genomes from the calculation and resulting in a significance score higher than it should be.

It is important to frame conclusions derived from the data of GenCoDB with the limitations from the data generation methods. Our definitions of neighbourhoods are based solely on gene order meaning factors such as intergenic distance, genetic elements such as repetitive sequences and non-coding RNAs are not included. The knowledge of intergenic distance can have important implications on genes for instance it has been shown that small distances (0-20bp) between genes transcribed unidirectionally are more likely to evolve into overlapping genes (where two transcriptional units share partial sequence) (Fukuda, Nakayama, and Tomita 2003). Overlapping genes are not displayed in GenCoDB and would have negative intergenic distances. Knowledge of intergenic distances is also crucial to many operon prediction algorithms where separation of less than 60 nucleotides between co-direction genes is sufficient to identify 75-80% of operons (Janga et al. 2006; Moreno-Hagelsieb and Collado-Vides 2002). As GenCoDB was built using ortholog groups defined by OrthoDB, which in turn utilizes similarities between protein amino acid sequences of encoded genes to categorize the groups, any transcripts which are not translated can therefore not be assigned an ortholog group nor included in our database. Non-coding RNAs are transcripts, usually 50-250 nucleotides in length (Eddy 2001) and can have very crucial and diverse roles including in processes like virulence (Toledo-Arana, Repoila, and Cossart 2007), stress response (Calderón et al. 2014) and quorum sensing (Bejerano-Sagie and Xavier 2007). They typically function by binding to mRNA resulting in either the stimulation/inhibition of translation or the degradation of stabilization of the mRNA. Therefore, they are similar to genes in that their conservation in a neighbourhood could also indicate relevance to the other genes/non-coding RNA in cluster. They are not as well annotated as translated genes because they lack open reading frames and small size makes them hard to detect on a sequence level. Thus, they are usually detected based on locating highly conserved intergenic regions (Wassarman et al. 2001) or areas which would form RNA secondary structures typically seen in non-coding RNAs such as stem-loops (Rivas et al. 2001). As GenCoDB utilizes NCBI genome annotations, both intergenic distances and non-coding features are accessible. The former could be implemented in future extensions of GenCoDB both in the genome view which could be scaled to display actual genomic proportions, and in the neighbourhood view, where hovering in the spaces between the bars could display a histogram showing the variation in intergenic distance between genes in these positions. The latter however will prove a challenge both without a unifying categorization system that can identify orthologous non-coding RNAs and the unequal levels of annotation between the genomes.

In order to determine both the level of conservation expected in different taxonomic groups and to reduce the impact genome sampling bias we utilized a genetic distance measure implemented in Vieira-Silva and Rocha, 2010. Here they used differences in 16S rRNA sequences as a proxy for genetic distance. The 16S rRNA gene is ubiquitous across bacteria and highly conserved with nine hypervariable regions that are used for species identification (Baker, Smith, and Cowan 2003). For this reason the 16S rRNA gene has been used to define bacterial taxonomy and the genetic distances between species (Mushegian and Koonin 1996). Utilization of the 16S rRNA sequences has limitations however. A threshold of 97% difference between the hypervariable regions has commonly been used to delineate bacterial species borders but this itself has been shown to be partially flawed. Two different species can have highly similar 16S rRNA sequences such as in the case of *Bacillus globisporus* and *Bacillus psychrophilus* with 99% similarity. Conversely, some genomes such as *E. coli* can have multiple 16S rRNA sequences which can differ by up to 5% (Eren et al. 2013). This is unsurprising as while evolution of the genome and the gene do correlate, they can have different histories, resulting in deviations between the evolutionary signals provided by them. In order to solve this issue and to acquire better resolution in the fine graining of species definitions, multiple well conserved genes are often used (such as elongation factors). However, the increase in the number of genes used also results in an increase in the complexity and computational time required. Therefore, as this method is for the resolution of evolutionary distances at the small scale, whereas GenCoDB focuses on long term evolution, we reasoned this was not worth the scale challenges it brings to future updates of the database. Another limitation of comparing 16S rRNA sequences is that it is an indirect measure of evolutionary time between species, and does not perfectly capture gene shuffling. We see that rate in which chromosomal rearrangements and shuffling events occur do vary in different clades sometimes independently of rRNA mutations as seen in figure 2.3. Aligning whole genomes is not thought to be a plausible solution and is complicated due to differing lengths of genomes, multiple chromosomes and lack of gene order. There have been other methods used to measure genetic distances between genomes that are not gene alignment based including gene content (Snel, Bork, and Huynen 1999), using the proteome to count the frequency of oligopeptide strings (Qi, Wang, and Hao 2004) and of relevance to genome organization the presence/absence of genes in found in conserved gene clusters (Wolf, Rogozin, Grishin, et al. 2001), gene order (House, Pellegrini, and Fitz-Gibbon 2014), and the persistence conserved gene pairs (Wolf, Rogozin, Grishin, et al. 2001). Unfortunately, either, these methods would not resolve the disconnect between the distance score and expected rearrangement events, or, in order to generate distance scores between thousands of genomes would require a much larger amount of computational time than is currently used. As methods develop and computational resources increase, it would be good to re-examine this limitation for GenCoDB and may help resolve the discrepancies we see between genetic distance and neighbourhood conservation in taxa such as Actinobacteria and Mollicutes.

Future updates to GenCoDB could bring numerous auxiliary functions which improve the usability of the web tool. An early expansion would be the inclusion of Archaea and eukaryotic genomes to the database. OrthoDB already includes these genomes and therefore the addition would not involve significant changes to the database. As “kingdom” is the highest offered taxonomic level by orthoDB it would be not possible to group together shared genes between both bacteria and eukaryotes. We outlined in the introduction how eukaryotes have different evolutionary mechanisms that shape their genome organization and therefore it would be interesting to observe the difference in neighbourhood statistics between both prokaryotes and eukaryotes. Archaeal genomes have often been included with bacteria in previous genome organization studies (Dandekar et al. 1998; Wolf, Rogozin, Kondrashov, et al. 2001) therefore including them in GenCoDB would be a valuable resource for archaea researchers and an extension of earlier work. The next change would be independence from OrthoDB. Currently GenCoDB is restricted to the update schedule of OrthoDB (approximately every 2 years) as it supplies the ortholog groupings required to perform the neighbourhood retrieval. It also limits the number of genomes we are able to include in the database, as they must be mapped by these ortholog group mappings, therefore the scalability and ability to handle large numbers of genome sequences of GenCoDB is not being used to its full extent. To achieve independence ortholog maps need to generated. To do this for many thousands of genomes is a complex and computationally taxing task; however, there are a few tools already published that would allow this with more development (Lechner et al. 2011; Camacho et al. 2009)

5.2 Neighbourhood Evolution Analysis

Using the over 1.9 million gene neighbourhoods we identified 1383 gene cluster families which were trackable over the evolutionary history of bacteria. By tracking the size of clusters through the different lineages we found no strong tendency in the expansion or degradation of gene clusters with only a slight bias towards clusters increasing in size (Figure 3.12). As there is not a large bias in the descendants of clusters this suggests our strategy to account for the increased conservation expected by evolutionary chance is satisfactory. There are two well accepted models on the size dynamics of gene clusters, the piecewise model and the uber-operon. The former states that gene clusters come together slowly overtime through stepwise building processes (Fani, Brilli, and Liò 2005). Fani et al. 2005 show evidence suggesting the progression of the histidine biosynthesis gene cluster in proteobacteria occurring in a stepwise manner. Whilst in our dataset we also observe the growth of some gene clusters this does appear to be an outlying behaviour and therefore should not be a model explaining gene cluster size dynamics generally. The piecewise model also predicts that expansion of the cluster should occur downstream as to minimize disruption of gene regulation (Fani, Brilli, and Liò 2005). However, we could see no bias to the localization of new genes to gene clusters (Figure 3.13). The uber-operon hypothesis states that operons splitting is a normal evolutionary event and once split the former genes remain in a similar functional and regulatory context (Lathe, Snel, and Bork 2000). As we do not see a bias towards shrinking clusters the frequency of this occurring must be limited, however we should temper our observation due to the methods used in cluster lineage detection. We considered a cluster a descendant if it contains at least 50% of the genes from an ancestor cluster. Therefore, if a cluster was fragmented into more than thirds, we no longer tracked the cluster and its shrinking lineage would not contribute to our statistics. The possibility also exists that both the piecewise model and uber-operon-based degradation occur simultaneously in close to equal levels resulting an observed average static size that we see. Given that cluster size seems unchanging, this is what would be expected if gene clusters were transmitted via HGT instead of vertical transmission. The selfish operon model posits that genes cluster together so that they are more likely to be transferred together in random HGT events (J. G. Lawrence and Roth 1996). By both normalizing cluster identification by either the taxonomic divisions or the conservation of the gene and comparing them, we were able to detect clusters which were present scattered across the bacteria kingdom. However, compared to the 1383 gene families identified with the former method, we found only 59 that could be attributed to HGT. Therefore, we believe HGT is not an important method in long term gene cluster formation. This is in agreement with Pál and Hurst 2004, who characterized HCT events by measuring the absence of essential genes (genes that are unlikely to be involved with HGT) from of gene clusters. As our method is independent of essentiality status, therefore including the possibility of essential genes transferring, we believe that we provide more weight behind this and therefore posit that vertical gene transmission is the underlying method of gene cluster evolution.

We studied the impact of the four main hypothesized selective pressures that could act in maintaining gene clusters: essentiality, intra-cluster interactivity, operon transcription, and transcriptional co-regulation. Co-regulation was found very rarely enriched in gene clusters (Figure 3.16) and this was especially true in clusters which have withstood longer periods of evolution (Figure 3.17). This is in agreement with the work of J. Lawrence, 1999 and J. G. Lawrence and Roth 1996, whom argue that the selective advantage of co-transcription can be matched more easily by stochastic evolutionary events resulting in regulons. The other three selective factors were enriched in relatively high abundances in gene clusters with our numbers closely agreeing with previous studies using significantly reduced sample sizes (Fang, Rocha, and Danchin 2008; Huynen et al. 2000). We measured higher levels of interacting proteins in our gene clusters than Huynen et al. 2000 found in their study comparing *Mycoplasma* species. We believe this difference lies in both their small sample size of only three genomes, and the improvement in our knowledge of protein-protein interactions since 2000, increasing the number of genes which are classified as interacting. Essential genes have previously been shown to be overrepresented in operons and gene clusters (Pál and Hurst 2004) and are usually found at the 5' end of a cluster (Muro et al. 2011). Conversely, Fang, Rocha, and Danchin 2008 state that it is not essentiality but persistence which is relevant to gene clustering. They define persistent genes as those which are highly conserved across the bacterial kingdom. There is significant overlap between persistent genes and essential genes, and indeed we found that there were several highly conserved genes with strongly conserved gene neighbourhoods (Figure 3.1); however, these did mainly consist of also genes we would classify as essential. To corroborate the two works, we believe it is most likely that genes, which provide strong fitness benefits cluster together, which is why they are either broadly conserved or classified as essential. However, the fitness benefit of every gene in every species is hard to quantify, especially outside of a laboratory setting, therefore we believe essentiality is a suitable substitute.

We saw that whilst several gene clusters were enriched in either interacting or essential genes, there were very few where operon organization was the only selective force (Figure 3.16). Again, this was most apparent in clusters which have been maintained over long periods of evolutionary history. We postulate that this means as a selective force, operon level transcription is not strong enough to resist the mutational pressure of genomic rearrangements. Therefore, the small increase in correlation provided by operon level transcription is most likely insufficient to overcome this barrier. In fact, it was shown by (Lathe, Snel, and Bork 2000) that operons regularly broken up into smaller operons which continue to be similarly regulated (referred to as uber-operons). The lack of importance of operons in gene clusters certainly has implications in our understanding of gene clusters. A conserved genomic context is often used as evidence of operon expression or to predict the presence of operons (Price et al. 2005). However, if operons are not important in maintaining gene clusters, why then would there have been an enrichment in polycistronic transcripts at the same level as the other factors? We would outline two possibilities to explain this observation. Firstly, is that operons are a fitness benefit, albeit weak in the face of mutational pressure. We outlined how polycistronic transcription benefits bacteria in the introduction, but briefly, co-transcription results in the transcript level of the transcribed genes to be in a 1:1 stoichiometry which has been shown to reduce expression noise (Ray and Igoshin 2012); furthermore, it reduces gene network complexity, improves the efficiency of RNA polymerase and requires the production of less transcription factors to control the same number of genes. Due to other factors (namely the essentiality or interactions of the genes) maintaining the gene cluster, the fitness benefit from operons is able to act and optimize expression of the cluster genes. It has been argued however, that the benefit of stoichiometry in transcripts is often not exploited by bacterial evolution. Many well studied operons have translational efficiency variations between different genes on polycistronic transcripts thereby resulting in the proteins no longer being in stoichiometry (J. Lawrence 1999). The second possibility is that it is the natural state for genes which are co-localized is to deteriorate into an operon co-transcribed state due to mutations of the internal terminators and transcription start sites. Previous work has suggested that operons develop through a piecewise model, where the size of the operon slowly increases over evolutionary time, i.e. operon before gene cluster (Fani, Brilli, and Liò 2005). This is counter what we postulate, as we believe our data suggests that the gene cluster comes first. However, the number of genes belonging to conserved gene clusters is relatively low (n=286 in *B. subtilis*) whereas the number of genes in multigene operons is relatively high (n=2504) therefore it is unlikely the model we postulate is a major driver in operon formation. We also observed that in conserved neighbourhoods and gene clusters, co-orientation is also highly conserved. Co-orientation is most likely one of the first factors that is resolved during the formation of a gene cluster, as we have shown within gene cluster rearrangements happen very infrequently (Figure 3.4). But which forces can explain the formation of co-orientation hand in hand with cluster formation, if the most obvious need for co-orientation (operons) is only established later? We would suggest that the directionality we observe is not a cause of operons but the negative consequences that would result from divergent or convergent gene orientation. Firstly, genes on bacterial chromosomes are usually expressed on the leading strand to reduce head-on collisions occurring between transcription and DNA replication machineries which has been shown to be detrimental to the cell (Paul et al. 2013). Additionally, in highly compressed genomes, transcriptional collision has been implicated in impairing transcription termination and elongation between convergent gene pairs (Prescott and Proudfoot 2002).

The majority of our evolutionary analysis focused on the clusters that still persist in the modern day *B.subtilis* genome. Whilst genes that are essential, in one organism are not essential in another, we show that genes which are highly conserved (and therefore often essential for life) have conserved neighbours (Figure 3.1) and therefore when studying the bacterial wide clusters our observations should hold. For clusters which were detected at lower taxonomic ranks (e.g Firmicutes, Bacilli) these likely are either not found in other taxonomic groups or behave differently. To test if our observations hold in other taxonomic contexts, future work should look at the extant clusters of species belonging to other distant taxonomic groups (such as *E.coli*). A further interesting avenue for research would be the correlation of genome size with the maintenance of gene clusters. As observed in taxa known to have average genome size differences, the strength of gene neighbourhood conservation also varied accordingly (Figure 2.3)

5.3 TCE Cluster

During changes in nutrient availability and subsequently through evolution, cells have learnt to adapt their physiology to optimize their growth rate requiring a balance between DNA replication, division and protein synthesis. How the cell manages to synchronize the increased demand of cell envelope biosynthesis to accommodate the increased volume growth seen at faster growth rates currently remains a mystery. We postulated this could be achieved through connecting the expression of key cell envelope intermediates with ribosome-associated genes. In our work in chapters 2 and 3, we developed the infrastructure allowing us to scan genomic neighbourhoods looking for this link connecting surface and volume growth. To our surprise, genes involved in cell envelope synthesis were very rarely found in context with translation associated genes (Table 4.1). We found only one cluster which connected three ribosomal-associated genes with cell envelope genes that was found well conserved in the bacterial kingdom (Table 4.1). This cluster consisted in total of eight genes: *rpsB*, *tsf*, *pyrH*, *frr*, *uppS*, *cdsA*, *dxr* and *rasP*. The mix of biological roles of the genes contrasts previous findings showing that gene clusters normally consist of genes involved in similar cellular processes (Fang, Rocha, and Danchin 2008; Rogozin et al. 2002). As well as the translation and cell envelope genes, there are also genes involved in pyrimidine metabolism and division. Earlier work that identified gene clusters in bacteria only identified parts of this cluster usually as gene pairs (namly *rpsB-tsf* and *uppS-cdsA*) (Rogozin et al. 2002). As these studies were limited at the time in available genomes, we believe our larger dataset and therefore higher resolution allowed us to detect this gene cluster.

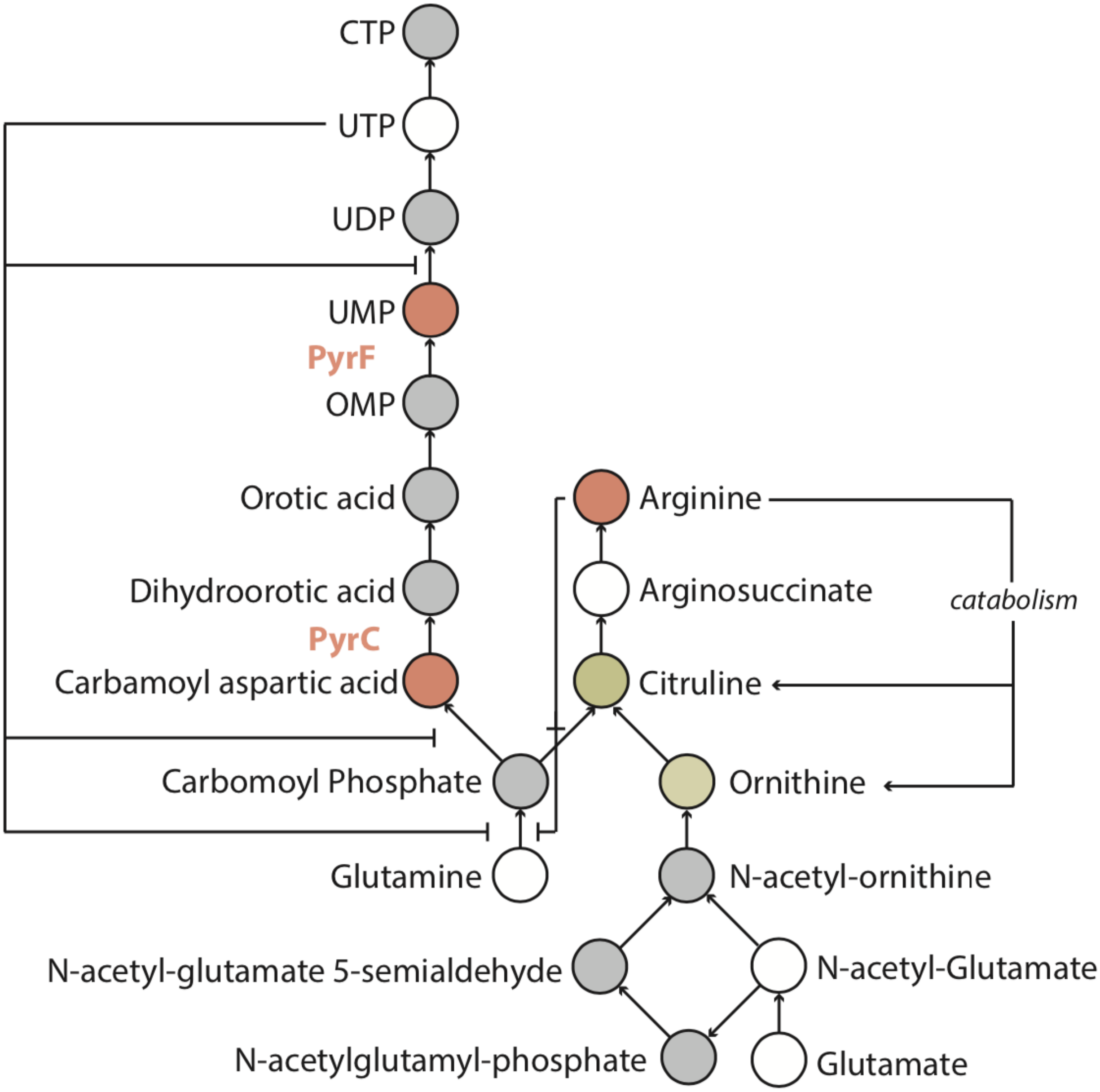
As both co-localization and co-orientation of these eight genes was highly conserved (Figure 4.23) it seemed likely that volume and surface growth could be synchronized through the co-regulation of these genes. This was supported when we realized that nearly all genes held essential bottleneck positions in their respective processes (Figure 4.2), and in organisms where duplication events made this untrue (*uppS* in actinobacteria) the gene was removed from the cluster (Figure 4.23). We speculated that perhaps by regulating expression of this locus, the rate of translation, cell wall synthesis and division could be controlled acting as a cellular clock. In order to confirm if these genes in fact transcriptionally correlated, we analysed public transcriptomic datasets from different species, and in different conditions. We saw high levels of co-regulation in all species where the cluster was maintained (Figure 4.6). However, we subsequently showed that the ribosome-associated genes and cell envelope biosynthesis genes was not transcribed on a single transcript. Therefore, these genes do not form an operon, despite previous evidence stating otherwise (Figure 4.8). Earlier in the discussion, we mentioned the possibility that operons naturally form over time in genes that are co-localized. As this cluster is in the context of evolution very old and must have formed in an early ancestor of both Firmicutes and Proteobacteria it is unlikely that lack of polycistronic transcripts is due to lack of evolutionary time. The reason may be linked to what we observed in the relative levels of expression of the eight genes in the different species where each gene was expressed at different levels in different species (Figure 4.7). By using CRISPRi we were able to perturb expression of each of the genes (Figure 4.13 and 4.14). In response to inhibition, the majority of the genes had different effects on the growth rate to the repression, contrary to what would be expected if every gene was rate limiting to growth. From these data, we cannot know for certain that each gene responded to the repression identically, and it should be noted that the genes do vary by several orders of magnitude of expression naturally. The work of Peters et al. 2016 did show similar repression from CRISPRi for different genes. If synchronization of these eight genes and their respective processes is essential, we would suggest co-regulation appears to play a minor, if ever present role.

In chapters 3 and 5.2, we discussed how co-regulation and operon organization is insufficient to maintain gene clustering (Figure 3.15). This suggests there is another selective pressure maintaining the TCE gene cluster. We explored three avenues: modulating size control with growth rate, increased enzyme efficiency due to transertion, and genomic channelling of pyrimidine metabolites.

Due to containing all the elements required to be able to regulate the documented cell size increase in response to growth rate (namely growth rate sensing regulation, cell wall synthesis and division control) we hypothesized the TCE cluster is responsible for this relationship. *rasP* was thought to be a key member in this hypothesized control as it cleaves a late stage division protein and in its absence was shown to result in a mini-cell phenotype (Bramkamp et al. 2006). This finding was only tested at one growth rate (Bramkamp et al. 2006) and we hypothesized that without *rasP* we would observe no concurrent change in cell size with the modulation of the growth. With our independent *rasP* deletion mutant we could confirm the mini cell phenotype from Bramkamp et al. 2006 however we found that this was only true in LB media (and presumably other fast growing media) (Figure 4.22). When grown in minimal media supplemented by non-preferred carbon sources of *B.subtilis* we found the mutant cells were actually larger than the wild type (Figure 4.22). This highlights the importance of testing phenotypes at multiple different growth rates. Whilst we did not see a complete loss in the relationship between growth rate and cell size, the deletion of *rasP* did lessen the strength of the relationship (Figure 4.22). This suggests that *rasP* is responsible for size control however it is not the sole contributor and there must be other proteins responsible for creating redundancy in cell size control. Previously this relationship has been perturbed by directly targeting cytoskeletal proteins to shrink the cell without changing the growth rate (Monds et al. 2014), or ribosomal maturation to slow growth rate without modulating size (Bügl et al. 2000). *rasP* is not directly involved with either as was the case with UgtP which normally acts in cell envelope biogenesis but moonlights as a metabolic sensor to link cell size with central metabolism (Weart et al. 2007; Hill et al. 2013). A deletion in *ugtP* results in a similar phenotype as the *rasP* mutant in that the relationship between growth rate and size was perturbed (Hill et al. 2013). The results in Hill et al. 2013 do not show the effect on size from this deletion in conditions slower than a doubling time ~50 minutes which is where the intersect occurs in our data between the wild type and *rasP* mutant (Figure 4.22). As we currently do not know what triggers RasP activity in the cell, we could postulate that RasP may also act as a metabolic sensor together with UgtP and that these proteins collectively ensure cell size and growth rate to be balanced. Aside from FtsL, RasP is known to cleave RsiV and RsiW, anti-sigma factors to their respective sigma factors σ^V and σ^W (Zweers et al. 2012). Our preliminary data showed that deletion of *sigW* but not *sigV* (therefore replicating what may happen to the concentration of σ^W and σ^V levels in the absence of *rasP*) resulted in a similar but weaker change in the size-growth rate relationship. This suggests the phenotype we observe by deleting *rasP* maybe a cumulative effect from the increase in FtsL and RsiW (and subsequently the reduction in σ^W) levels. As we were able to restore the growth rate relationship by complementing *rasP* at an independent locus (Figure 4.22), this suggests that the genomic context of *rasP* associating it with ribosome-associated genes and expression is not essential. To further understand the role of *rasP* in the size-growth rate relationship we modulated the expression of *rasP* by changing the inducer concentrations. This would confirm if the relationship is the result of RasP concentration levels and may indicate the dynamic range in which *rasP* functions. An interesting observation is that in the slope of the relationship between size and growth rate for the wild type, *rasP*, *sigW* and *ugtP* deletion mutants, was that despite their slopes all being different they all intersected at approximately a doubling time of 50 minutes. Therefore, at a doubling time faster than every 50 minutes wild type cells were larger than the mutants and vice versa. The question is why is this 50-minute doubling time conserved? It is thought that a reason for size expansion during increasing growth rates is to accommodate the bulk of DNA arising from multifork replication. Perhaps not coincidentally the time it takes for *B.subtilis* to replicate its genome is 40-50 minutes (Skarstad and Katayama 2013). This however raises the question, why would the cell want to be smaller during slow growth conditions (> 50-minute doubling time)? Smaller cells have a higher surface area to volume ratio increasing the ability for external nutrients to sustain the requirements of the cell which may be necessary in poor conditions that generate these slow growth rates. Further still, smaller bacteria have a height dry weight per volume and it is suggested the reduction of water content may reduce energetic costs (Simon and Azam 1989). Therefore, in *B.subtilis*, the 50 minute doubling time might represent a breakpoint where the advantages of a larger cell volume to accommodate multifork replication outweigh the advantages of being small. We should note that despite the observed size changes in all these mutations, we did not see a significant change in growth rate (Figure 4.22). Suggesting either the difference in size is accommodated for by the cell, or that there are other physiological changes that do not perturb the growth rate.

Within the TCE cluster is uridylate kinase (*pyrH*) an essential enzyme in the synthesis of pyrimidine producing UDP from UMP (Figure 5.1). It is co-localized alongside three enzymes which either directly (*cdsA*) or indirectly through their downstream reactions (*uppS, dxr*) require UDP or the downstream product CTP (Figure 4.2). Additionally, by looking at which genes correlated in genomes where the cluster was maintained we found that in the absence of a cluster both arginine decarboxylase and uridine kinase were both enriched (Table 4.3). Species which have functional uridine kinase are less reliant on *de novo* pyrimidine synthesis as they can scavenge UMP from uridine. Similarly, species expressing arginine decarboxylase may have more abundant sources of arginine (or a more efficient *de novo* biosynthesis pathway) and therefore more readily catabolize it. Therefore, this suggests that the selective pressures maintaining TCE cluster may be highly interwoven with a functional and well-regulated pyrimidine and arginine pathway. The molarity model suggests a selective advantage of gene clustering to be the increased local concentration of proteins resulting in more efficient protein-protein interactions and metabolic channelling of substrates. Therefore, *pyrH*'s position in the chromosome may facilitate the use of its product UDP by other pyrimidine consumers downstream. We reasoned by separating the pyrimidine producer from the consumers would result in the deregulation of the pyrimidine metabolites and subsequently arginine metabolism due to the shared intermediates (Figure 5.1). By genetically splitting the cluster (Figure 4.25) we indeed found that whilst there were no strong negative phenotypes in rich media, a strong slow growth phenotype in MOPS minimal media was observed in the absence of arginine (Figure 4.27).

To explain this observation, we have two hypotheses. Firstly, due to the inefficient utilization of pyrimidines by pyrimidine consumers normally found in the cluster, there is an accumulation of UDP and other pyrimidines. Pyrimidines are known to be negative regulators of the branch molecule between arginine and pyrimidine metabolism, carbamoyl phosphate, thereby resulting in decreased flux and production of arginine, and arginine deprivation for the cell (Turner, Lu, and Switzer 1994). By supplementing arginine, the requirement for arginine *de novo* synthesis from carbamoyl phosphate is no longer needed resulting in normal growth. Secondly, *pyrH* is known to be localized to the membrane (Noria and Danchin 2002); however, the biological significance of this has not yet been elucidated. It is believed the compartmentalization may separate it from other enzymes that would readily convert UDP to dUPD potentially resulting in the production of the deleterious dTMP. It may also be degraded away from dUMP to prevent its conversion to dUDP resulting in the formation of a futile cycle. Perhaps the loss of co-localization with cell envelope genes is important for the localization of PyrH. This delocalization would result in decreased efficiency of PyrH potentially limiting the cells of pyrimidines. Preliminary data showed that of the four antibiotics tested, which all targeted different cellular processes, only rifampicin was shown to be more effective in the split cluster mutant in the arginine deprived conditions potentially referencing the lack of pyrimidine nucleotides for transcription (Figure 4.29). With the addition of arginine, all flux from carbamoyl phosphate could be directed to pyrimidine biosynthesis, potentially counteracting the loss in efficiency. When analysing UTP levels in the split cluster we did not see a significant difference in UTP concentration however it is known that *pyrH* is autoregulated by UTP and therefore it is unsurprising we see no difference (Turner, Lu, and Switzer 1994) (Figure 5.1). Unfortunately, the utilized method for measuring metabolites was unable to measure UDP levels at this stage. We do see that two other metabolites in the pyrimidine pathway are reduced, namely UMP and carbamoyl aspartic acid, which could suggest that flux into this pathway is reduced.



66 - Figure 5.1 - The effect of TCE de-localisation on Pyrimidine and arginine metabolism

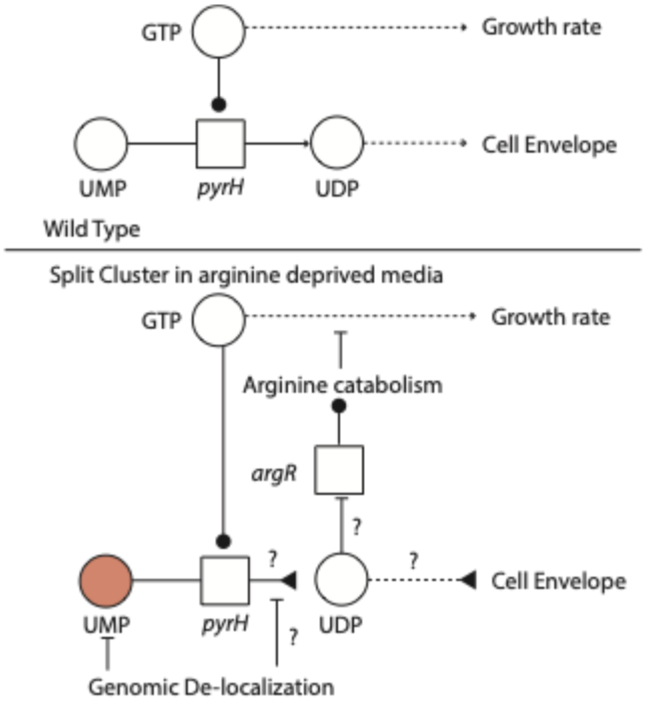
A schematic of the pyrimidine and arginine *de novo* biosynthesis pathways. Circles represent metabolites. The shade represents its status in the metabolomics data set: white - unchanged, red - reduced concentration in the split cluster mutant, green - increased concentration in the split cluster mutant, grey - unmeasured. Proteins involved in this pathway that were significantly differentially expressed are written in bold between the metabolites with the same colour scheme. Lines with a blunt end represent negative regulation from the metabolite acting on the step in the pathway.

Strikingly, metabolomic data revealed that arginine metabolism was perturbed. Even in media where arginine was supplemented the split cluster mutant had close to no intracellular arginine, similar to the levels found in the wild type and split cluster in minimal media (Figure 4.34). We do not see an upregulation of arginine-rich proteins which suggests the arginine is not being sequestered away by translation. This leaves two other possibilities, that the split cluster mutant is exporting/not importing arginine or they are consuming it an increased rate. As we are able to restore the wild type phenotype with the supplementation the latter seems more likely. In bacteria there are many arginine catabolic pathways, three main ones include: the arginase pathway, the ADI pathway, and the arginine succinyltransferase pathway, reviewed in (C.-D. Lu 2006). The arginase pathway has arginine catabolized into ornithine by arginase, releasing urea, ornithine is then covered to glutamate by ornithine aminotransferase which can be converted further into 2-ketoglutarate resulting in a source of carbon and nitrogen for the cell (Gardan, Rapoport, and Débarbouillé 1995). In bacteria where urease is present, urea may also act as a nitrogen source. The arginine deiminase (ADI) pathway creates energy, carbon and nitrogen sources for the cell. Arginine deiminase converts arginine to both L-citrulline and ammonia, the former then being broken down further into ornithine and carbamoyl phosphate (Broman et al. 1978). Carbamoyl phosphate can be used to produce pyrimidines or is further broken down into ammonia, CO_2 and forms ATP by carbamate kinase. It is thought that this pathway is a main supplier of ATP during anaerobic conditions (Noens and Lolkema 2017). The arginine succinyltransferase pathway also utilizes arginine and ornithine as a carbon and nitrogen source resulting in 2 molecules of ammonia and 2 glutamate and is induced during carbon starvation conditions (Cunin et al. 1986). This pathway is mainly found in proteobacteria (Stalon et al. 1987) and has been shown to be the main cause of arginine degradation in *E. coli (Schneider, Kiupakis, and Reitzer 1998)*. In some species, this catabolic pathway allows arginine to be the sole carbon or nitrogen source for the cell (Cunin et al. 1986). Based on metabolite changes in our split cluster samples we saw that citruline and orinithine pool levels were higher, and in the presence of arginine ATP levels were much higher (Figure 4.31). This would suggest that the arginine diminase pathway is currently active however this pathway has so far not been detected in the *B. subtilis* W168 strain that we use.

The proteomic dataset revealed that protein levels for the arginine repressor (ArgR) were lower in the split cluster mutant. ArgR repressor activity is regulated by the concentration of arginine in the cell and controls 423 genes in *E .coli (Cho et al. 2011)*. When arginine is present in the cell, ArgR both activates arginine metabolism genes and represses biosynthetic genes (Czaplewski et al. 1992). ArgR is auto-regulated, therefore at high levels of cellular arginine ArgR levels are normally low and vice versa (Tian et al. 1994). Given the low levels of arginine in the mutant, one would expect high levels of ArgR in the split cluster. however, this is not the case. (Table 4.5). Lower levels ArgR would result in less repression of arginine catabolism genes and could explain while even in low-arginine conditions we see metabolic evidence that catabolism is occurring. Notably, knockouts of ArgR were shown to result in reduced growth rates (Sander et al. 2019); however, this is attributed to the many other targets AgrR has. The work of Sander et al. 2019 revealed that deletion or repression of ArgR activity non-uniformly affected the genes in the arginine metabolism pathway with ArgA and ArgI being most affected. ArgI catalyzes the branch point reaction between arginine and pyrimidine biosynthesis and thereby reduce the available metabolites for the pyrimidine pathway. The proteomics method was unable to detect peptides from the core arginine metabolism genes; therefore, we cannot determine if this is occurring in our system. However, we do see an increase in citrulline which would be expected if repression of ArgI is relieved by lower repressor concentrations. (Sander et al. 2019) were able to partially restore the growth rate by supplying pyrimidine intermediate orotate and the precursor aspartate. In agreement with their results we see large drops in carbamoyl aspartate and orotic acid (Figure 4.31) but we do not see significant reductions in UTP levels as they did. In their work however the drop in UTP levels was not as large as the other two metabolites suggesting that their reduction mainly has consequences on the metabolites prior to UTP (namely in our case UMP). The reduction in UMP levels in the split cluster mutant may explain why in species which do not localize the gene cluster, the UMP salvaging enzyme uridine kinase is present.

Glutamate is the precursor for arginine biosynthesis but acts as the end product of arginine catabolism and can enter the TCA cycle through glutamine dehydrogenase producing α-ketoglutarate. Expression of glutamate dehydrogenases is normally activated by arginine (not glutamate) and inhibited by citrate. Arginine also represses glutamate synthesis. Through this control *P. aeruginosa* can control the flow of glutamate into the tricarboxylic acid cycle to prevent the futile and energy consuming cycle in ammonium assimilation and glutamate biosynthesis when arginine serves as a source of carbon and nitrogen. While neither glutamate nor glutamate dehydrogenase changes in the split cluster mutant we see an increase in α-ketoglutarate despite a reduction in isocitrate dehydrogenase (Figure 4.31). Glutamate dehydrogenase is inhibited by citrate, which we were unable to measure, however citrate synthase was shown to be upregulated suggesting that citrate synthesis is up. (Table 4.5). This could suggest that arginine is being broken down into α-ketoglutarate resulting either in the observed increase in ATP or a futile cycle, wasting the resources of the cell and resulting in the slow growth phenotype. To further understand what is occurring the split cluster, we would propose supplying labelled arginine to the media to determine exactly which metabolic products are using arginine. This may help us understand what is triggering arginine metabolism if it is in fact occurring. Canonically, arginine catabolism is triggered during high intracellular levels of arginine (Gardan, Rapoport, and Débarbouillé 1995). However it has already been found to be triggered by ornithine, citruline and proline, and mediated through σ^L (Gardan, Rapoport, and Débarbouillé 1995). The former two metabolites are present in higher levels in the split cluster (Figure 4.31), and presumably citrulline levels could be increased if flux from carbamoyl phosphate is directed to arginine biosynthesis due to negative regulation of *pyrBI* by over accumulating UDP.

Taken together, with our incomplete metabolic and proteomic datasets, it is hard to rule out our hypothesis as to why we observe slow growth in the split cluster mutant. Parsimoniously, the culprit appears to be limited arginine levels in the cell caused by unregulated arginine catabolism, however the link as to why the removal of co-localization of the TCE cluster would cause this remains elusive. Given the tight regulatory links between pyrimidines and arginine biosynthesis (Sander et al. 2019), *pyrH* or one its substrates seems like the likely cause. Therefore, we propose a model based on the information we have generated so far, that delocalized *pyrH*, from pyrimidine consumers (*uppS, cdsA, dxr)* result in a disruption of *de novo* pyrimidine biosynthesis, namely UMP, carbamoyl aspartic acid and potentially UDP by directing flux of carbamoyl phosphate towards arginine biosynthesis, as evidenced by the higher levels of citrulline (Figure 5.2). This potentially occurs through an unknown mechanism that downregulates the arginine repressor independent of arginine concentration resulting in the induction of arginine catabolism. Arginine is then catabolized resulting in either more citrulline or into the carbon cycle producing ATP. Subsequently, this induces a state of arginine starvation in arginine lacking media, as any new synthesized arginine is metabolised (Figure 5.2). Given the ever-present context of the ribosome-associated genes we must ask if the coordination of pyrimidine metabolites through *pyrH* and its consumers is related to growth rate. Our analysis of transcription over the TCE cluster does not preclude the existence of a *pyrH-frr* operon therefore UDP production could be statistically linked with the rate of translation. Further support is that PyrH is activated by GTP levels. GTP has been shown to control ribosomal promoter activity in a growth dependant manner (Gaal et al. 1997) thereby linking growth rate with the production of UDP. UDP could then be used to modulate the synthesis of cell envelope genes and the link between carbon metabolism, UDP-glucose modules and cell division (Hill et al. 2013). This leaves us with the following questions (Figure 5.2): through what mechanism is *argR* downregulated, where is the arginine going, is the relationship between GTP levels and PyrH important for cell growth, and are there effects on the cell envelope that our assays were not able to detect? One possible avenue is to test how the delocalization functions in an *argR* depleted background. *argR* is unessential and can be deleted, with growth rate defects in certain conditions, however if it is the sole cause of the slow growth phenotype, we should see no difference between a split and non-split mutant.



67 - Figure 5.2 - Model of TCE cluster delocalization on growth in minimal media

In the wild type PyrH is responsible for the production of UDP which is then used in many aspects of the cell, one being the cell envelope, which has many important enzymes co-localized with pyrH. If the TCE cluster is split, thereby separating the cell envelope pyrimidine consumers from the producer we see effects in pyrimidine metabolism (red). We also see argR downregulation resulting in the activation of inappropriate arginine catabolism resulting in arginine starvation in arginine poor conditions. We postulate that the dysregulation of pyrimidines is caused by the separation of the consumers from the producer and through an unknown mechanism, downregulates *argR*.

Our split cluster mutant is built so that that the latter four genes (*upps*, *cdsA*, *dxr* and *rasP*) are translocated to the *amyE* locus (Figure 4.25). We provided every effort to try and control collateral effects from this manipulation, for example maintaining a similar level of expression of the downstream essential genes (*proS* and *polC*) (Figure 4.28), and ensuring the slow growth phenotype in minimal media was not caused by the artificial expression either *uppS* or *proS* (Figure 4.28). However, we should understand the limitations of our observations. Firstly, the new locus of *amyE* is considerably closer to the origin of replication which means that at different growth rates there will be gene dosage effects (Soler-Bistué, Timmermans, and Mazel 2017) resulting in increased expression of these four genes at high growth rates. *B. subtilis* requires 40-50 minutes to replicate its chromosome (Skarstad and Katayama 2013) and therefore would not need to undergo multifork replication in MOPS media (multifork replication normally occurs once the doubling time reaches 35 minutes in rich media) therefore we should only observe these effects in the LB media condition. However, as gene dosage effects are predicted to be relevant in ribosome-growth rate regulation as the ribosome supercluster is also located near the origin (Soler-Bistué, Timmermans, and Mazel 2017) we suggest further controls placing the latter four genes at other loci in the genome to remove this as a factor. Another limitation is we currently observe only the effects caused by a split of the cluster occurring between *ffr* and *uppS.* This location was chosen for several reasons. Firstly, we wanted to understand if this cluster connected surface and volume expansion and between these two genes marks the demarcation between the three translation associated genes and the cell envelope genes. Secondly, as we aimed to understand the fitness benefit of the co-localization of the eight cluster genes, by moving half of the cluster, this results in the highest combination of genes being separated from each other which we hoped would have the largest impact on fitness. Finally *uppS* is the first gene in the cluster which has levels of expression mimicable by the inducible promoters currently available in *B.subtilis* (Radeck et al. 2013). In our attempts to clone the promoter of *rpsB* in *B.subtilis* we found it was lethal in our cloning system, most likely due to the high affinity promoter in multiple copies titrating RNA polymerase away from the essential transcription required in the cell. If these challenges could be overcome, based on the data we presented highlighting the perturbation of pyrimidine metabolism in the split cluster (Figure 4.31) we would suggest that further splits between *pyrH* and *frr* or *uppS* and *cdsA*, thereby still separating the pyrimidine producer from the majority of the consumers would result in a similar phenotype. We believe that splits that maintain context between *pyrH* and the pyrimidine consumers would not result in the slow growth phenotype in minimal media. This is supported as the minimal translocation of *rasP* (Figure 4.22) did not result in any noticeable differences in the growth rate in various media conditions.

5.4 Conclusions

The aim of this study was to search for and analyse conserved genomic signals which could indicate a method of synchronization between cell surface and volume growth required when adapting to different growth rates. To this end we developed the web-tool GenCoDB, a platform that facilitates quantitative and statistical analysis of bacterial gene neighbourhoods. We developed this tool to handle the rapidly expanding numbers of bacterial genomes. It is designed to reduce sampling bias present in bacterial genome datasets and calculate the significance of observed gene clusters (Chapters 2.1 and 2.2). The webtool comes with three different forms of user interfaces (neighbourhood, tree and genome view) enabling diverse types of analyses and research questions (Chapter 2.3). Using the data from GenCoDB we analysed the evolution of gene neighbourhoods. We found, in agreement with previous work, the enrichment of operons in conserved gene clusters (Chapter 3.3). However, careful evolutionary analysis revealed their role as a selective force was weak and was insufficient to explain gene cluster maintenance alone. We proposed a model suggesting that operons are a consequence of gene clusters and not cause suggesting observed gene clusters are present for reasons beyond the synchrony of their gene products (Chapter 3.3). We applied both GenCoDB and our new knowledge of gene cluster evolution to the search for a genomic context involved in the synchronization of surface and volume growth. The genomic co-localisation of cell envelope synthesis genes with ribosomal-associated genes was rare with only one non-canonical candidate being found (Chapter 4.1). The TCE cluster was identified to be a well conserved gene cluster found mainly in proteobacteria, firmicutes and actinobacteria and consisting of mainly of essential rate-limiting enzymes (Chapter 4.2). In agreement with our observations from Chapter 3, we found that polycistronic transcripts did not link the expression translation and cell envelope genes. However, subsequent delocalization of the gene cluster in *B. subtilis* suggested that cluster localization is essential for appropriate pyrimidine biosynthesis and utilization (Chapters 4.8 to 4.11). Future work can continue to unravel the selective pressures maintaining the TCE cluster and its possible role in synchronizing volume and surface growth.

6. Materials and Methods

6.1 Lab-Bench methods

Chemicals and enzymes

All chemicals, enzymes and enzyme buffers used were purchased from Carl Roth GmbH & Co. KG (Karlsruhe, Germany), Thermo Scientific (Waltham, Massachusetts), Sigma- Aldrich (Saint Louis, Missouri), AppliChem (Chicago, Illinois), New England Biolabs (NEB) (Ipswich, Massachusetts) or BD (Franklin Lakes, New Jersey), if not mentioned otherwise.

Strains and growth conditions

Bacillus subtilis and Escherichia coli were routinely grown in Luria-Bertani (LB) medium (1% (w/v) tryptone, 0.5% (w/v) yeast extract, 1% (w/v) NaCl) at 37°C with agitation (250rpm). Solid media additionally contained 1.5% (w/v) agar. All strains used are shown in Table 8.1. Selective media for *B. subtilis* contained spectinomycin (100µg/ml), chloramphenicol (5µg/ml), erythromycin in combination with lincomycin (1µg/ml: 25µg/ml for mlsr). Selective media for *E. coli* contained IPTG (0.1mM) and Xgal (40µg/ml) with either ampicillin (100µg/ml) or spectinomycin (50µg/ml). For the growth assays cells were grown overnight in 3ml and day cultured in 10ml (1:250 dilution from the overnight culture) in MOPS minimal media (10% 10X MOPS mixture (Teknova, Hollister, CA), 1% K2HPO4, 88% sterile H2O). MOPS media was either supplemented with only L-Tryptophan (0.25µg/ml) or a mixture of L-Methionine, L-Histidine, L-Arginine, L-Proline, L-Threonine and L-Tryptophan at the same concentration. When not specified, glucose was added as the carbon source (1.8% w/v). Other carbon sources included: xylose (1% w/v), fructose (1.8% w/v), glycerol (1.6% w/v), ribose (0.8% w/v) and succinate (1% w/v).

Creation of Level 0 parts

The majority of plasmids we generated using a modular cloning method (MoClo) (Weber et al. 2011). All plasmids generated using this method and used in this study can be found in tables 8.2-8.5. Table 8.6 that indicates the primers and the templates utilized to generate the part inserts (by PCR-amplification or oligonucleotides annealing), together with the MoClo destination vectors used for each part. The constructs that required to be cured for BpiI and BsaI restriction sites are also indicated with multiple forward and reverse primers. The cure of undesired BpiI and BsaI sites was performed according to Weber *et al*. To generate the genetic parts present in the library we used PCR-amplification or annealing of DNA oligonucleotides. In the case of PCR-amplification, the PCR products were verified by electrophoresis with 1, or 2% agarose gels and purified by gel extraction or column purification, following the protocols of the manufacturer. The purified product was used to clone the insert into the appropriate MoClo destination vector, following the procedure described in a later section. In case of annealing of DNA oligonucleotides, the reaction of annealing and the phosphorylation of the 5'OH was performed as follows: 2 µL of 100 µM oligonucleotides stock were mixed with 2 µL 10X T4 DNA ligase buffer, 1 µL of T4 Polynucleotide Kinase and 15 µL of sterile water. The reaction mixture was incubated at 37 °C for 1 hour and at 65 °C for 20 minutes to heat inactivate the T4 PNK. An aliquot of reaction mix was then used to clone the insert into the appropriate MoClo destination vector.

Modular Cloning (MoClo) reactions (Golden gate assembly).

All constructs were assembled in MoClo, using linear DNA fragments (PCR-amplificated products, or phosphorylated annealed oligonucleotides) or the MoClo-encoded parts listed in Table

8.2 (level 0 parts), Table 8.3 (level 1 parts), Table 8.3 (level M parts). The parts each vector is constructed on are displayed as a combination of ID numbers which reference a part in another table (e.g Level M parts (TABLE X) references level 1 parts (TABLE X). Each table indicates the list of the parts used to generate the constructs and a brief description of the constructs. All MoClo reactions were set up using 15 fmol of each DNA part (PCR product or plasmid), 1μL of the required restriction enzyme (BsaI or BpiI), 1 μL of T4DNA ligase (5 U/μL) and 2 μL of Thermo ligase buffer (10x), in a final reaction volume of 20 μL. The reaction was incubated in a thermocycler for 5 h at 37 °C, 10 min at 50 °C and 10 min at 80 °C. 2 μL of the reaction mixture was then added to 50 μL chemically competent *E. coli* DH5α cells (*E. coli* DH5α *λpir* cells in case of CRIMoClo constructs), incubated for 30 min on ice and transformed by heat shock. 950μL of liquid LB was then added to the transformation, and the cells were recovered for 45 min at 37 °C. 40 μL of the transformation mix was plated on selective LB-IPTG-X-Gal plates and incubated overnight at 37 °C. The emerging colonies were tested by colony PCR and restriction digestion.

Creation of other plasmids

Plasmids that were generated not using the MoClo method can be found in Table X. To create the CRISPRi we followed the protocol from (Peters et al. 2016). sgRNA sequences for the TCE genes were when possible also taken from (Peters et al. 2016) as they were experimentally tested and verified. the sgRNA for rasP was designed using Bowtie, and the highest scoring 19nt (rasP -based) sequence which had only one alignment was chosen, with a preference to those at the 5' end of the gene. Sequences can be found in table 8.6. These primers were used with GF0561 to inverse PCR amplify pJMP2 and then the plasmid was relegated to later be transformed.

Transformation of *E. coli* strains

E. coli DH5α cells were inoculated in 125ml SOB from a 5μl overnight culture. The culture was then incubated for 15-17 hours at 22C with shaking (120rpm), until an OD of 0.5 was reached. The cultures were then put on ice for 10 minutes, spun at 2500xg for 10 minutes at 4C. Cells were then re-suspended in 40ml of chilled TB. Put on ice again and process repeated, this time re-suspended in 10ml of TB. Finally, 0.7ml of DMSO added and put on ice for 10 minutes. Cells were then aliquoted in Eppendorf tubes (into 100μl), and snap frozen in liquid nitrogen and stored at -80C. Transformations of the E. coli DH5α strains were carried out according to the standard protocol. The competent cells were defrosted on ice for 20 minutes, DNA added (5μl ligation mixture per 50μl cells) and mixed, left for 30 minutes before 50 second heat shock at 42°C. After a 2-minute cooling period on ice, LB was added to give a total volume of 1ml and then placed into a shaking incubator at 37°C for one and a half hours. Cells were then plated onto LB agar plates with selective antibiotics and IPTG / Xgal and incubated at 37°C overnight. Spectinomycin resistance plates used for level 0 and level M and ampicillin resistance plates used for level 1.

Transformation of *B. subtilis* strains

Plasmids for *B. subtilis* transformation were prepared from the *E. coli* overnight cultures containing the desired plasmid and purified using the Omega E.Z.N.A. plasmid DNA mini kit. The plasmid was linearised. This linearized plasmid was used for transformation without further purification. A day before the transformation B. subtilis was streaked out on a LB agar plate to create a bacterial lawn. The incubation was done overnight at 30C. In the morning Medium 1 and Medium 2 were prepared. Medium 1 was inoculated with the number of bacteria on an agar plate that was needed to reach an OD of 0.2. This solution was incubated for 3 hours at 37 C while shaking. Afterwards

10 ml pre-warmed medium 2 was added and the culture for another 2 hours incubated. After that time 400 µl of the culture was transferred to a test tube and 5 ul linearized DNA was added. Cells were then plated onto LB agar plates with the appropriate antibiotic.

Basic Salts	**Medium 1**	**Medium 2**
2.0g/l (NH4)2SO4	10ml Basic Salts	10ml Basic Salts
14.0g/l K2HPO4	120µl 40%(w/v) Glucose	120µl 40%(w/v) Glucose
6.0g/l KH2PO4	100µl 2%(w/v) Tryptophan	60µl 1M MgSO4 x 7H2O
1.0g/l Na3-citrate x 2H20	60µl 1M MgSO4 x 7H2O	
0.2g/l MgSO4 x 7H2O	10µl 20%(w/v) Casaminoacids	
	5µl 2.2mg/ml Ferric-ammoniumcitrate	

RNA purification

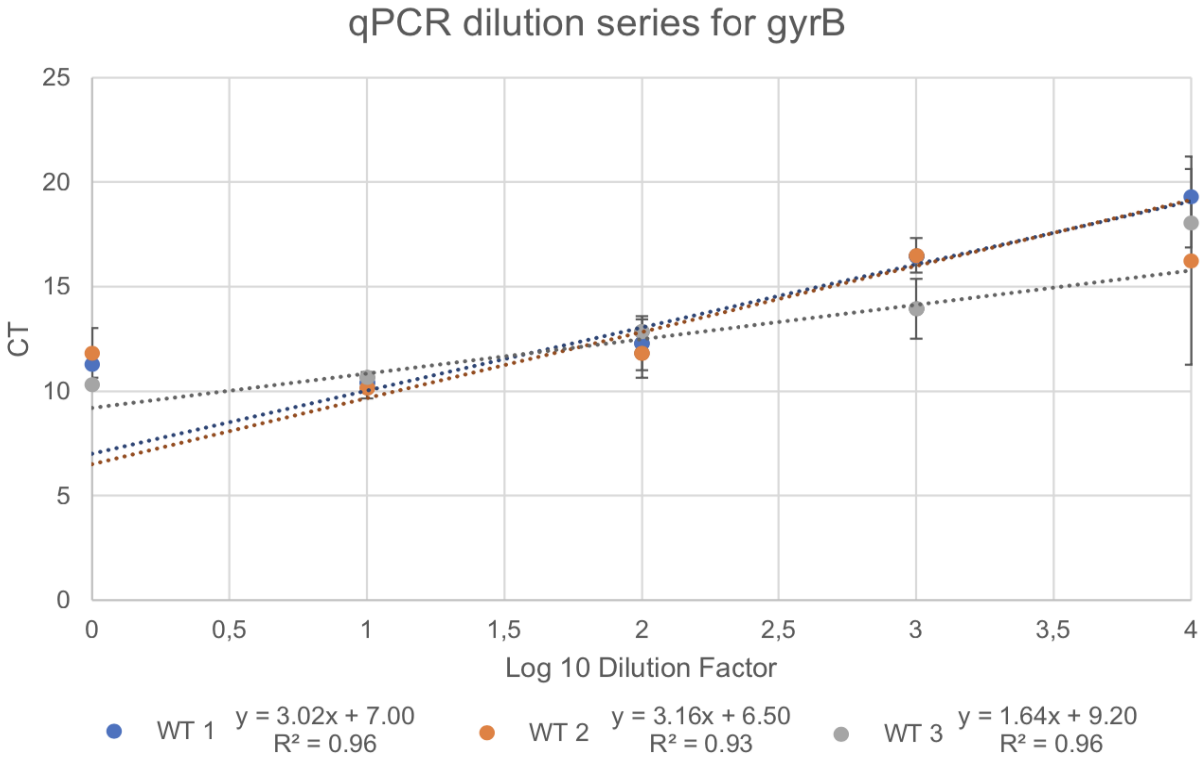
Bacillus subtilis cells (20ml) were harvested at an optical density (OD600) between 0.3 and 0.5 via centrifugation (10 minutes, 5000rpm, room temperature) in the specified media. Supernatant was removed and 1.5mls of TRIzol Reagent (Ambion) was added. The resuspended cells were combined with 0.1 zirconia beads and lysed in a bead beater (3 cycles, 6.0 m/s, time 40seconds, 5min pause). RNA isolation was performed using the standard TRIzol reagent protocol.

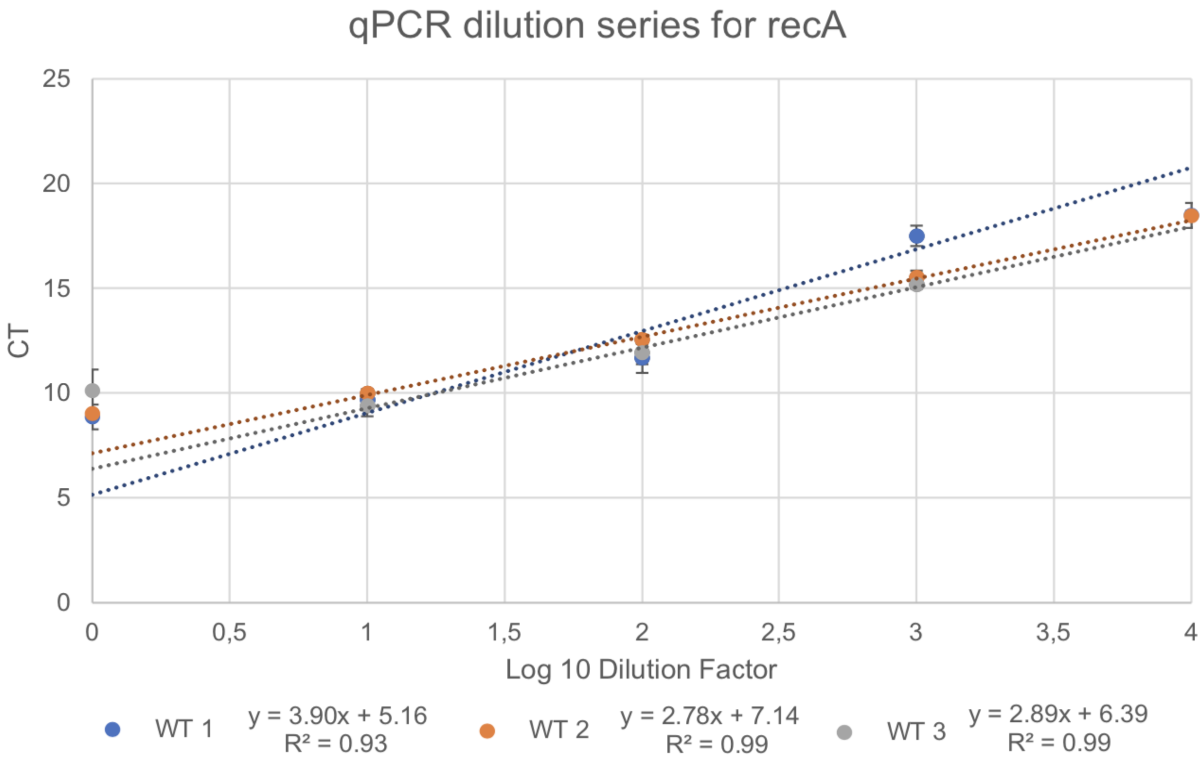
5'RACE

~0.5ug RNA was used to generate cDNA using the standard NEB reverse transcriptase protocol. A cocktail mixture of gene specific primers was used to prime the reverse transcriptase. The cDNA was then purified and then A-tailed using the standard NEB TdT protocol. After another purification step the cDNA was amplified using an A-tail anchor primer (5' GACCACGCGTATCGATGTCGACTTTTTTTTTTTTTTTTC 3') and a gene specific nested primer (94C, 15s denaturation, 50C 30s annealing, 72C 40s elongation, x35). The PCR product was purified and another nested PCR was performed using the adapter primer (5' GACCACGCGTATCGATGTCGAC 3') and another gene specific nested primer (94C, 15s denaturation, 59C 30s annealing, 72C 40s elongation, x35). This PCR product was visualized on a gel, gel extracted if muliple bands were present, and send for sequencing using the nested primer. All 50ul PCR reaction included 1.25ul of DMSO to reduce secondary structures.

qPCR

RT-qPCR was performed with the Luna Universal One-Step RT-qPCR Kit (New England Biolabs) on extracted RNA. 1 µl of 10-fold diluted RNA was added to 4 µl of rtPCR mix and subjected to a reverse transcription step at 55°C and 45 cycles of PCR (10) seconds at 95°C and 30 seconds at 60°C). The average CT value of three technical replicates of three biological replicates for each sample was used in ΔΔCt relative expression analysis (Livak and Schmittgen, 2001). The reference genes were constitutively expressed genes recA (BSU16940) and gyrB (BSU00060) (da Silva et al. 2016; Crawford et al. 2014; Gomes et al. 2018; Reiter, Kolstø, and Piehler 2011). Primer sequences can be found in table 8.6. In figure 6.1 we show using our two reference genes that the 10-fold dilution of all six RNA samples fell in the linear detection range for qPCR and that no dilution or a dilution greater than 10,000 would have resulted in some of the samples being incorrectly measured.


qPCR dilution series for gyrB
25
20
15
10
5
0
CT
0
0,5
1
1,5
2
2,5
3
3,5
4
Log 10 Dilution Factor
● WT 1
y = 3.02x + 7.00
R² = 0.96
● WT 2
y = 3.16x + 6.50
R² = 0.93
● WT 3
y = 1.64x + 9.20
R² = 0.96


qPCR dilution series for recA
25
20
15
10
5
0
CT
0
0,5
1
1,5
2
2,5
3
3,5
4
Log 10 Dilution Factor
● WT 1
y = 3.90x + 5.16
R² = 0.93
● WT 2
y = 2.78x + 7.14
R² = 0.99
● WT 3
y = 2.89x + 6.39
R² = 0.99

68 - Figure 6.1 – qPCR dilution series for reference genes

Each dot represents the average of 3 technical replicates at a different dilution of the RNA sample. A dilution factor of 0 represents no dilution of the extracted RNA. CT is Cycle Threshold. The linear regression was fit only to the points with a log 10 dilution factor of 1 – 3 as some samples fall out of the linear range when not diluted or diluted too much. The equation for the slope and R^2 value can be found for each line in the legend.

Plate Reader assays

Cells were grown overnight in 3ml culture of either LB or MOPS minimal media. A 10ml day culture of MOPS minimal media was inoculated with 40µl of the overnight culture to give a 1:250 dilution. This was grown to an OD of 0.1-0.2 and then diluted to give the same OD of 0.05. 100µl of these dilutions were added to the 96-well plate (Grainer 655097), for control purposes some wells were also filled with MOPS minimal media. For the induction with xylose or bacitracin, 5µl of water and xylose or bacitracin at varying concentrations were added to the samples. The final xylose concentrations were 0-2% and the final bacitracin concentrations 0-100µg/µl. Upon induction the plate reading began, shaking at 37°C and the OD of each well was measured in real time every 10 minutes for hours. As a background control, wells were filled with MOPS minimal media only. The plate was inserted into the plate reader (Victor2). The analysis, calculations and visual representations were made using MATLAB. The average OD600 values of the control wells (containing only minimal media) were averaged and used as a blank, in order to remove background noise from the reads.

Isolation of genomic DNA from *B. subtilis*

For isolation of genomic DNA from Bacillus subtilis a 3 ml LB culture was done overnight at 37C. In the morning 10 ml LB medium were inoculated from the overnight culture (1000-fold dilution). At OD600 of 0.8 – 1.0 the culture was centrifuged to harvest the cells (10 min, 5000 rpm, RT). The pellet was resuspended in 400µl TEN and transferred into 2 ml eppendorf cups. Then we added 20 µl lysozyme and incubate for 20 min at 37ÅãC. After the incubation 2µl RNase A was added and incubated for another 3 min at 65C. In the next step 40 µl SDS, a small amount (covering a tip of a small spatula) of proteinase K and 550 µl TEN* were added and mixed and Incubated for 2 hours at 60C. Then 900 µl of phenol (equilibrated with TE buffer, pH 7,5-8.0) were added and the solution mixed by inverting the tube. In the next step, the tubes were centrifuged (5 min, 130 000 rpm, RT) and the upper phase was transferred into a new 1.5 ml eppendorf cup. The extraction was repeated once with phenol and twice with chloroform: isoamyl alcohol (24:1). The aqueous phase was then transferred to 10 ml -20C cold ethanol in a test tube / falcon tube. The precipitated DNA can then be coiled up with a bent tip of
a Pasteur pipettes. DNA was air tried and dissolved in TEN* or ddH2O overnight at 4C.

Microscopy

To prepare slides for microscopy experiments cells were cultured overnight and freshly inoculated in 3ml media for day culture. At an OD600 of ~0.3 -0.5 cells were harvested. 0.5 ml of the cell culture was shortly spun down and 200 µl of the supernatant removed. The Cells were placed on 1% MOPS media agar pads. Viewed under a phase microscope. The 514nm laser was used to visualize YFP activity. Cell counting, measuring and loci tracking were performed in oufti (Paintdakhi et al. 2016).

6.2 Bioinformatics

Transcriptomic dataset analysis

RNAseq data were downloaded from the NCBI SRA archive, all datasets used can be found in table 8.9. The raw RNA-seq data were processed with fastq-mcf to remove sequencing adapters and primers (Aronesty 2011). The reads were quality trimmed to a Phred score of >20 using SolexaQA v3.1.4 (Cox, Peterson, and Biggs 2010). Analysis of the general quality parameters of the raw and processed data was done using FastQC v0.11.5 ("Babraham Bioinformatics - FastQC A Quality Control Tool for High Throughput Sequence Data" n.d.). The reads were mapped to the respective reference genome using Bowtie 2 (version 2.2.6) with --sensitive and --end-to-end (Langmead and Salzberg 2012). Where possible paired end reads were used. The reads were then assigned to the gene features from the respective gff files with HTSeq-count using union mode on exon features (Anders, Pyl, and Huber 2014; Li et al. 2009). Mapping were visualized in the Integrative Genome viewer (Thorvaldsdóttir, Robinson, and Mesirov 2013). Correlation analysis was performed with R. Conservation analysis was performed with a personalized python script.

Data collection

Flat files were downloaded from OrthoDB v10 (Kriventseva et al. 2015). An archived version of the UniProt ID mapping (Jan 2019) to match the data from OrthoDB . GFF files were acquired from NCBI as of 1 Jan 2019.16S rRNA sequences of the included species were retrieved from SILVA resource (Quast et al. 2013). Taxonomy was used as defined by NCBI taxonomy. Further analysis was performed on the strains Bacillus subtilis W168, Escherichia Coli K12 and Mycobacterium tuberculosis H37Rv. Mapping to orthoDB equivalents was performed by taking the top BlastP hit (v2.7.1 with default parameters) to Bacillus subtilis subsp. natto BEST195 (645657), Escherichia coli TW10509 (656449) and Mycobacterium tuberculosis CDC1551 (83331) respectively (Camacho et al. 2009). Operon membership was sourced from DOOR (Database of Prokaryotic Operons) (v2.0) (Mao et al. 2009). Essentialness of genes were defined by the DEG database (v15.2) (Luo et al. 2014). Protein-protein interactions were retrieved from the STRING database (v10.5) (Snel et al. 2000). For the analysis of genomic location of clusters, the origin of replication locations were acquired from OriDB (v2.1.0) (Siow et al. 2012). 16S rRNA sequences were aligned using Clustal Omega (v1.2.3) using the default parameters (Sievers et al. 2011). A tree was built on this alignment using Fasttree (v2.1.10) with the default parameters (Price, Dehal, and Arkin 2010). Pairwise patristic distances of each species in the tree was calculated by PATRISTIC (v1.0) (Fourment and Gibbs 2006). To discriminate between which ortholog groups would be classified as the conserved synteny at each taxonomic level the most conserved groups surrounding the seed at each position were identified. If the positional conservation of that group is higher than the threshold it was considered part of a part of the conserved synteny. This process radiated out from the seed gene until there are no more significantly conserved groups. In the case a middle position did not have a group above the significance threshold it is displayed as a blank.

6.3 Omics

DNA sequencing

Several split cluster mutant strains were allowed to grow in MOPS media without amino supplementation. Cultures which survived and reached stationary phase were inoculated into the same media conditions and tested for wild type growth rate. One candidate was chosen was streaked out for single colonies. Reinoculation and DNA extraction of the single colony was performed as described above. Extracted DNA was sent for library preparation and sequencing to AG Becker. The raw data sequencing data was treated the same as the RNAseq data sets. SNPs were detected using the Genome Analysis Toolkit (McKenna et al. 2010) and their effect on genes was calculated by SNPEff (Cingolani et al. 2012).

Metabolomics

Cultures of wild type and split cluster were inoculated into MOPS media with amino supplementation. They were allowed to grow until an optical density (OD 600) of 0.3 and they were spun down and washed with MOPS media (no amino acid supplementation). The cultures were divided and amino acids were re-added to one half. The cultures were then allowed to regrow to and OD of 0.5. For metabolomics 2 mL culture aliquots were vacuum-filtered on a 0.45 µm pore size filter (HVLP02500, Merck Millipore). Filters were immediately transferred into 40:40:20 (v-%) acetonitrile/methanol/water at −20 °C for extraction. Extracts were centrifuged for 15 min at 11,000 × g at −9 °C. Centrifuged extracts were mixed with 13C-labeled internal standard. Chromatographic separations were performed on an Agilent 1290 Infinity II LC System (Agilent Technologies) equipped with an Acquity UPLC BEH Amide column (2.1 × 30 mm, particle size 1.8 µm, Waters) for acidic conditions and an iHilic-Fusion (P) HPLC column (2.1 × 50 mm, particle size 5 µm, Hilicon) for basic conditions. We were applying the following binary gradients at a flow rate of 400 µl min−1: acidic condition) 0–1.3 min: isocratic 10% A (water/formic acid, 99.9/0.1 (v/v), 10 mM ammonium formate), 90% B (acetonitrile/formic acid, 99.9/0.1 (v/v)); 1.3–1.5 min linear from 90 to 40% B; 1.5–1.7 min linear from 40 to 90% B, 1.7–2 min isocratic 90% B. Basic condition) 0–1.3 min: isocratic 10% A (water/formic acid, 99.8/0.2 (v/v), 10 mM ammonium carbonate), 90% B (acetonitrile); 1.3–1.5 min linear from 90 to 40% B; 1.5–1.7 min linear from 40 to 90% B, 1.7–2 min isocratic 90% B. The injection volume was 3.0 µl (full loop injection). Eluting compounds were detected using an Agilent 6495 triple quadrupole mass spectrometer (Agilent Technologies) equipped with an Agilent Jet Stream electrospray ion source in positive and negative ion mode. Source gas temperature was set to 200 °C, with 14 L min−1 drying gas and a nebulizer pressure of 24 psi. Sheath gas temperature was set to 300 °C and flow to 11 L min−1. Electrospray nozzle and capillary voltages were set to 500 and 2500 V, respectively. Metabolites were identified by multiple reaction monitoring (MRM), and MRM parameters were optimized and validated with authentic standards44. Metabolites were measured in 12C− and 13C isoforms, and data were analyzed with Metabolanalyst (Chong, Wishart, and Xia 2019).

Proteomics

Test-tube cultivations on MOPS media with amino acid supplementation were performed as described above for wild type and split cluster mutant strains. Cells were grown to an optical density (OD600) of 0.5- and 2-mL culture aliquots were transferred into 2 mL reaction tubes and washed two times with PBS buffer (0.14 mM NaCl, 2.7 mM KCL, 1.5 KH2PO4, 8.1 Na2HPO4). Cell pellets were resuspended in 300 µL of lysis buffer containing 100 mM ammonium bicarbonate, 0.5% sodium laroyl sarcosinate (SLS), and 5 mM Tris(2-carboxyethyl)phosphine (TCEP). Cells were lysed by 5 min incubation at 95 °C and ultrasonication for 10 s (Vial Tweeter, Hielscher). Cells were again incubated for 30 min at 90 °C followed by alkylation with 10 mM iodoacetamide for 30 min at 25 °C. To clear the cell lysate, samples were centrifuged for 10 min at 15 000 rpm, and the supernatant was transferred into a new tube. Proteins in the cell lysates were digested with 1 µg of trypsin (Promega) overnight at 30 °C. The analysis of peptides was performed by liquid chromatography–mass spectrometry, carried out on a Q-Exactive Plus instrument connected to an Ultimate 3000 RSLC Nano with a Prowflow upgrade and a nanospray flex ion source (Thermo Scientific). Peptide separation was performed on a reverse-phase HPLC column (75 µm × 42 cm) packed in-house with C18 resin (2.4 µm, Dr. Maisch GmbH, Germany). The following separating gradient was used: 98% solvent A (0.15% formic acid) and 2% solvent B (99.85 acetonitrile, 0.15% formic acid) to 25% solvent B over 105 min and to 35% solvent B for additional 35 min at a flow rate of 300 nL min–1. The data acquisition mode was set to obtain one high resolution MS scan at a resolution of 70 000 full width at half-maximum (at m/z 200) followed by MS/MS scans of the 10 most intense ions. To increase the efficiency of the MS/MS attempts, the charged state screening modus was enabled to exclude unassigned and singly charged ions. The dynamic exclusion duration was set to 30 s. The ion accumulation time was set to 50 ms for MS and 50 ms at 17 500 resolution for MS/MS. The automatic gain control was set to 3 × 106 for MS survey scans and 1 × 105 for MS/MS scans. Label-free quantification (LFQ) of the data was performed using Progenesis QIP (Waters), and for MS/MS searches of aligned peptide features MASCOT (v2.5, Matrix Science) was used. The following search parameters were used: full tryptic search with two missed cleavage sites, 10 ppm MS1 and 0.02 Da fragment ion tolerance. Carbamidomethylation (C) as fixed, oxidation (M), and deamidation (N,Q) as variable modification. Progenesis outputs were further processed with SafeQuant.

7. Bibliography

Bibliography

Achaz, Guillaume, Eric Coissac, Pierre Netter, and Eduardo P. C. Rocha. 2003. "Associations between Inverted Repeats and the Structural Evolution of Bacterial Genomes." *Genetics* 164 (4): 1279–89.

Aldén, L., F. Demoling, and E. Bååth. 2001. "Rapid Method of Determining Factors Limiting Bacterial Growth in Soil." *Applied and Environmental Microbiology* 67 (4): 1830–38.

Aldridge, Bree B., Marta Fernandez-Suarez, Danielle Heller, Vijay Ambravaneswaran, Daniel Irimia, Mehmet Toner, and Sarah M. Fortune. 2012. "Asymmetry and Aging of Mycobacterial Cells Lead to Variable Growth and Antibiotic Susceptibility." *Science* 335 (6064): 100–104.

Amir, Ariel. 2014. "Cell Size Regulation in Bacteria." *Physical Review Letters* 112 (20): 208102.

Anders, S., P. T. Pyl, and W. Huber. 2014. "HTSeq--a Python Framework to Work with High-Throughput Sequencing Data." *Bioinformatics* , September. https://doi.org/10.1093/bioinformatics/btu638.

Aronesty, Erik. 2011. "Ea-Utils: Command-Line Tools for Processing Biological Sequencing Data."

"Babraham Bioinformatics - FastQC A Quality Control Tool for High Throughput Sequence Data." n.d. Accessed July 27, 2015. http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/.

Baker, G. C., J. J. Smith, and D. A. Cowan. 2003. "Review and Re-Analysis of Domain-Specific 16S Primers." *Journal of Microbiological Methods* 55 (3): 541–55.

Balleza, Enrique, Lucia N. López-Bojorquez, Agustino Martínez-Antonio, Osbaldo Resendis-Antonio, Irma Lozada-Chávez, Yalbi I. Balderas-Martínez, Sergio Encarnación, and Julio Collado-Vides. 2009. "Regulation by Transcription Factors in Bacteria: Beyond Description." *FEMS Microbiology Reviews* 33 (1): 133–51.

Barka, Essaid Ait, Parul Vatsa, Lisa Sanchez, Nathalie Gaveau-Vaillant, Cedric Jacquard, Jan P. Meier-Kolthoff, Hans-Peter Klenk, Christophe Clément, Yder Ouhdouch, and Gilles P. van Wezel. 2016. "Taxonomy, Physiology, and Natural Products of Actinobacteria." *Microbiology and Molecular Biology Reviews: MMBR* 80 (1): 1–43.

Bejerano-Sagie, Michal, and Karina Bivar Xavier. 2007. "The Role of Small RNAs in Quorum Sensing." *Current Opinion in Microbiology* 10 (2): 189–98.

Biebricher, C. K., and M. Druminski. 1980. "Inhibition of RNA Polymerase Activity by the Escherichia Coli Protein Biosynthesis Elongation Factor Ts." *Proceedings of the National Academy of Sciences of the United States of America* 77 (2): 866–69.

Bonner, E. R., J. N. D'Elia, B. K. Billips, and R. L. Switzer. 2001. "Molecular Recognition of Pyr mRNA by the Bacillus Subtilis Attenuation Regulatory Protein PyrR." *Nucleic Acids Research* 29 (23): 4851–65.

Borkowski, Olivier, Anne Goelzer, Marc Schaffer, Magali Calabre, Ulrike Mäder, Stéphane Aymerich, Matthieu Jules, and Vincent Fromion. 2016. "Translation Elicits a Growth Rate-Dependent, Genome-Wide, Differential Protein Production in Bacillus Subtilis." *Molecular Systems Biology* 12 (5): 870.

Bouhss, A., D. Mengin-Lecreulx, D. Blanot, J. van Heijenoort, and C. Parquet. 1997. "Invariant Amino Acids in the Mur Peptide Synthetases of Bacterial Peptidoglycan Synthesis and Their Modification by Site-Directed Mutagenesis in the UDP-MurNAc:L-Alanine Ligase from Escherichia Coli." *Biochemistry* 36 (39): 11556–63.

Bramkamp, Marc, Louise Weston, Richard A. Daniel, and Jeff Errington. 2006. "Regulated Intramembrane Proteolysis of FtsL Protein and the Control of Cell Division in Bacillus Subtilis." *Molecular Microbiology* 62 (2): 580–91.

Bremer, Hans, and Patrick P. Dennis. 2008. "Modulation of Chemical Composition and Other Parameters of the Cell at Different Exponential Growth Rates." *EcoSal Plus* 3 (1). https://doi.org/10.1128/ecosal.5.2.3.

Broman, K., N. Lauwers, V. Stalon, and J. M. Wiame. 1978. "Oxygen and Nitrate in Utilization by Bacillus Licheniformis of the Arginase and Arginine Deiminase Routes of Arginine Catabolism and Other Factors Affecting Their Syntheses." *Journal of Bacteriology* 135 (3): 920–27.

Brown, Stephanie, John P. Santa Maria Jr, and Suzanne Walker. 2013. "Wall Teichoic Acids of Gram-Positive Bacteria." *Annual Review of Microbiology* 67: 313–36.

Bügl, H., E. B. Fauman, B. L. Staker, F. Zheng, S. R. Kushner, M. A. Saper, J. C. Bardwell, and U. Jakob. 2000. "RNA Methylation under Heat Shock Control." *Molecular Cell* 6 (2): 349–60.

Calderón, Iván L., Eduardo H. Morales, Bernardo Collao, Paulina F. Calderón, Catalina A. Chahuán, Lillian G. Acuña, Fernando Gil, and Claudia P. Saavedra. 2014. "Role of Salmonella Typhimurium Small RNAs RyhB-1 and RyhB-2 in the Oxidative Stress Response." *Research in Microbiology* 165 (1): 30–40.

Camacho, Christiam, George Coulouris, Vahram Avagyan, Ning Ma, Jason Papadopoulos, Kevin Bealer, and Thomas L. Madden. 2009. "BLAST+: Architecture and Applications." *BMC Bioinformatics* 10 (December): 421.

Campos, Manuel, Ivan V. Surovtsev, Setsu Kato, Ahmad Paintdakhi, Bruno Beltran, Sarah E. Ebmeier, and Christine Jacobs-Wagner. 2014. "A Constant Size Extension Drives Bacterial Cell Size Homeostasis." *Cell* 159 (6): 1433–46.

Cao, Min, Bryan A. Bernat, Zhepeng Wang, Richard N. Armstrong, and John D. Helmann. 2001. "FosB, a Cysteine-Dependent Fosfomycin Resistance Protein under the Control of ςW, an Extracytoplasmic-Function ς Factor in Bacillus Subtilis." *Journal of Bacteriology* 183 (7): 2380–83.

Chang, D. E., S. Shin, J. S. Rhee, and J. G. Pan. 1999. "Acetate Metabolism in a Pta Mutant of Escherichia Coli W3110: Importance of Maintaining Acetyl Coenzyme A Flux for Growth and Survival." *Journal of Bacteriology* 181 (21): 6656–63.

Chatterji, D., and A. K. Ojha. 2001. "Revisiting the Stringent Response, ppGpp and Starvation Signaling." *Current Opinion in Microbiology* 4 (2): 160–65.

Chen, I-Min A., Ken Chu, Krishna Palaniappan, Manoj Pillay, Anna Ratner, Jinghua Huang, Marcel Huntemann, et al. 2019. "IMG/M v.5.0: An Integrated Data Management and Comparative Analysis System for Microbial Genomes and Microbiomes." *Nucleic Acids Research* 47 (D1): D666–77.

Cho, Byung-Kwan, Stephen Federowicz, Young-Seoub Park, Karsten Zengler, and Bernhard Ø. Palsson. 2011. "Deciphering the Transcriptional Regulatory Logic of Amino Acid Metabolism." *Nature Chemical Biology* 8 (1): 65–71.

Chong, Jasmine, David S. Wishart, and Jianguo Xia. 2019. "Using MetaboAnalyst 4.0 for Comprehensive and Integrative Metabolomics Data Analysis." *Current Protocols in Bioinformatics / Editoral Board, Andreas D. Baxevanis ... [et Al.]* 68 (1): 375.

Cingolani, Pablo, Adrian Platts, Le Lily Wang, Melissa Coon, Tung Nguyen, Luan Wang, Susan J. Land, Xiangyi Lu, and Douglas M. Ruden. 2012. "A Program for Annotating and Predicting the Effects of Single Nucleotide Polymorphisms, SnpEff: SNPs in the Genome of Drosophila Melanogaster Strain w1118; Iso-2; Iso-3." *Fly* 6 (2): 80–92.

Cooper, S., and C. E. Helmstetter. 1968. "Chromosome Replication and the Division Cycle of Escherichia Coli B/r." *Journal of Molecular Biology* 31 (3): 519–40.

Couturier, Etienne, and Eduardo P. C. Rocha. 2006. "Replication-Associated Gene Dosage Effects Shape the Genomes of Fast-Growing Bacteria but Only for Transcription and Translation Genes." *Molecular Microbiology* 59 (5): 1506–18.

Cox, Murray P., Daniel A. Peterson, and Patrick J. Biggs. 2010. "SolexaQA: At-a-Glance Quality Assessment of Illumina Second-Generation Sequencing Data." *BMC Bioinformatics* 11 (September): 485.

Crawford, Evan C., Ameet Singh, Devon Metcalf, Thomas W. G. Gibson, and Scott J. Weese. 2014. "Identification of Appropriate Reference Genes for qPCR Studies in Staphylococcus Pseudintermedius and Preliminary Assessment of icaA Gene Expression in Biofilm-Embedded Bacteria." *BMC Research Notes* 7 (July): 451.

Cunin, R., N. Glansdorff, A. Piérard, and V. Stalon. 1986. "Biosynthesis and Metabolism of Arginine in Bacteria." *Microbiological Reviews* 50 (3): 314–52.

Czaplewski, L. G., A. K. North, M. C. Smith, S. Baumberg, and P. G. Stockley. 1992. "Purification and Initial Characterization of AhrC: The Regulator of Arginine Metabolism Genes in Bacillus Subtilis." *Molecular Microbiology* 6 (2): 267–75.

Dai, Xiongfeng, Zichu Shen, Yiheng Wang, and Manlu Zhu. 2018. "Sinorhizobium Meliloti, a Slow-Growing Bacterium, Exhibits Growth Rate Dependence of Cell Size under Nutrient Limitation." *mSphere* 3 (6). https://doi.org/10.1128/mSphere.00567-18.

Dandekar, T., B. Snel, M. Huynen, and P. Bork. 1998. “Conservation of Gene Order: A Fingerprint of Proteins That Physically Interact.” *Trends in Biochemical Sciences* 23 (9): 324–28.

Darling, Aaron E., István Miklós, and Mark A. Ragan. 2008. “Dynamics of Genome Rearrangement in Bacterial Populations.” *PLoS Genetics* 4 (7): e1000128.

Daubin, Vincent, Emmanuelle Lerat, and Guy Perrière. 2003. “The Source of Laterally Transferred Genes in Bacterial Genomes.” *Genome Biology* 4 (9): R57.

Dehal, Paramvir S., Marcin P. Joachimiak, Morgan N. Price, John T. Bates, Jason K. Baumohl, Dylan Chivian, Greg D. Friedland, et al. 2010. “MicrobesOnline: An Integrated Portal for Comparative and Functional Genomics.” *Nucleic Acids Research* 38 (Database issue): D396–400.

Derbyshire, Keith M., and Todd A. Gray. 2014. “Distributive Conjugal Transfer: New Insights into Horizontal Gene Transfer and Genetic Exchange in Mycobacteria.” *Microbiology Spectrum* 2 (1). https://doi.org/10.1128/microbiolspec.MGM2-0022-2013.

Dillingham, Mark S., and Stephen C. Kowalczykowski. 2008. “RecBCD Enzyme and the Repair of Double-Stranded DNA Breaks.” *Microbiology and Molecular Biology Reviews: MMBR* 72 (4): 642–71, Table of Contents.

Donachie, W. D. 1968. “Relationship between Cell Size and Time of Initiation of DNA Replication.” *Nature* 219 (5158): 1077–79.

Dong, Hengjiang, Lars Nilsson, and Charles G. Kurland. 1996. “Co-Variation of tRNA Abundance and Codon Usage inEscherichia Coliat Different Growth Rates.” *Journal of Molecular Biology* 260 (5): 649–63.

Dorman, Charles J. 2006. “DNA Supercoiling and Bacterial Gene Expression.” *Science Progress* 89 (Pt 3-4): 151–66.

Eddy, S. R. 2001. “Non-Coding RNA Genes and the Modern RNA World.” *Nature Reviews. Genetics* 2 (12): 919–29.

Eiamphungporn, Warawan, and John D. Helmann. 2008. “The Bacillus Subtilis sigma(M) Regulon and Its Contribution to Cell Envelope Stress Responses.” *Molecular Microbiology* 67 (4): 830–48.

El-Gebali, Sara, Jaina Mistry, Alex Bateman, Sean R. Eddy, Aurélien Luciani, Simon C. Potter, Matloob Qureshi, et al. 2019. “The Pfam Protein Families Database in 2019.” *Nucleic Acids Research* 47 (D1): D427–32.

Eren, A. Murat, Loïs Maignien, Woo Jun Sul, Leslie G. Murphy, Sharon L. Grim, Hilary G. Morrison, and Mitchell L. Sogin. 2013. “Oligotyping: Differentiating between Closely Related Microbial Taxa Using 16S rRNA Gene Data.” *Methods in Ecology and Evolution / British Ecological Society* 4 (12). https://doi.org/10.1111/2041-210X.12114.

Fadda, Daniela, Carla Pischedda, Fabrizio Caldara, Michael B. Whalen, Daniela Anderluzzi, Enrico Domenici, and Orietta Massidda. 2003. “Characterization of divIVA and Other Genes Located in the Chromosomal Region Downstream of the Dcw Cluster in Streptococcus Pneumoniae.” *Journal of Bacteriology* 185 (20): 6209–14.

Fang, Gang, Eduardo P. C. Rocha, and Antoine Danchin. 2008. “Persistence Drives Gene Clustering in Bacterial Genomes.” *BMC Genomics* 9 (January): 4.

Fani, Renato, Matteo Brilli, and Pietro Liò. 2005. “The Origin and Evolution of Operons: The Piecewise Building of the Proteobacterial Histidine Operon.” *Journal of Molecular Evolution* 60 (3): 378–90.

Fiil, N., and J. D. Friesen. 1968. “Isolation of ‘Relaxed’ Mutants of Escherichia Coli.” *Journal of Bacteriology* 95 (2): 729–31.

Fisher, R. A. 1929. *The Genetical Theory of Natural Selection*. Edinburgh: Oliver & Boyd.

Fondi, Marco, Giovanni Emiliani, and Renato Fani. 2009. “Origin and Evolution of Operons and Metabolic Pathways.” *Research in Microbiology* 160 (7): 502–12.

Fourment, Mathieu, and Mark J. Gibbs. 2006. “PATRISTIC: A Program for Calculating Patristic Distances and Graphically Comparing the Components of Genetic Change.” *BMC Evolutionary Biology* 6 (January): 1.

Francis, F., S. Ramirez-Arcos, H. Salimnia, C. Victor, and J. R. Dillon. 2000. “Organization and Transcription of the Division Cell Wall (dcw) Cluster in Neisseria Gonorrhoeae.” *Gene* 251 (2): 141–

51.

Fuente, A. de la, P. Palacios, and M. Vicente. 2001. “Transcription of the Escherichia Coli Dcw Cluster: Evidence for Distal Upstream Transcripts Being Involved in the Expression of the Downstream ftsZ Gene.” *Biochimie* 83 (1): 109–15.

Fukuda, Yoko, Yoichi Nakayama, and Masaru Tomita. 2003. “On Dynamics of Overlapping Genes in Bacterial Genomes.” *Gene* 323 (December): 181–87.

Gaal, T., M. S. Bartlett, W. Ross, C. L. Turnbough Jr, and R. L. Gourse. 1997. “Transcription Regulation by Initiating NTP Concentration: rRNA Synthesis in Bacteria.” *Science* 278 (5346): 2092–97.

Gagyi, Cristina, Nadia Bucurenci, Ovidiu Sîrbu, Gilles Labesse, Mihaela Ionescu, Augustin Ofiteru, Liliane Assairi, et al. 2003. “UMP Kinase from the Gram-Positive Bacterium Bacillus Subtilis Is Strongly Dependent on GTP for Optimal Activity.” *European Journal of Biochemistry / FEBS* 270 (15): 3196–3204.

Gagyi, Cristina, Mihaela Ionescu, Pierre Gounon, Hiroshi Sakamoto, Jean-Claude Rousselle, and Christine Laurent-Winter. 2004. “Identification and Immunochemical Location of UMP Kinase from Bacillus Subtilis.” *Current Microbiology* 48 (1): 62–67.

Ganong, B. R., J. M. Leonard, and C. R. Raetz. 1980. “Phosphatidic Acid Accumulation in the Membranes of Escherichia Coli Mutants Defective in CDP-Diglyceride Synthetase.” *The Journal of Biological Chemistry* 255 (4): 1623–29.

Gardan, R., G. Rapoport, and M. Débarbouillé. 1995. “Expression of the rocDEF Operon Involved in Arginine Catabolism in Bacillus Subtilis.” *Journal of Molecular Biology* 249 (5): 843–56.

Gärtner, D., J. Degenkolb, J. A. Ripperger, R. Allmansberger, and W. Hillen. 1992. “Regulation of the Bacillus Subtilis W23 Xylose Utilization Operon: Interaction of the Xyl Repressor with the Xyl Operator and the Inducer Xylose.” *Molecular & General Genetics: MGG* 232 (3): 415–22.

Geiman, Theresa M., and Keith D. Robertson. 2002. “Chromatin Remodeling, Histone Modifications, and DNA Methylation-How Does It All Fit Together?” *Journal of Cellular Biochemistry* 87 (2): 117–25.

Gerasimova, T. I., and V. G. Corces. 2001. “Chromatin Insulators and Boundaries: Effects on Transcription and Nuclear Organization.” *Annual Review of Genetics* 35: 193–208.

Gerlt, John A., Jason T. Bouvier, Daniel B. Davidson, Heidi J. Imker, Boris Sadkhin, David R. Slater, and Katie L. Whalen. 2015. “Enzyme Function Initiative-Enzyme Similarity Tool (EFI-EST): A Web Tool for Generating Protein Sequence Similarity Networks.” *Biochimica et Biophysica Acta* 1854 (8): 1019–37.

Ginez, Luis David, Aurora Osorio, and Sebastian Poggio. 2014. “Localization of the Outer Membrane Protein OmpA2 in Caulobacter Crescentus Depends on the Position of the Gene in the Chromosome.” *Journal of Bacteriology* 196 (15): 2889–2900.

Gomes, Ana Érika Inácio, Leonardo Prado Stuchi, Nathália Maria Gonçalves Siqueira, João Batista Henrique, Renato Vicentini, Marcelo Lima Ribeiro, Michelle Darrieux, and Lúcio Fábio Caldas Ferraz. 2018. “Selection and Validation of Reference Genes for Gene Expression Studies in Klebsiella Pneumoniae Using Reverse Transcription Quantitative Real-Time PCR.” *Scientific Reports* 8 (1): 9001.

Gómez, Manuel J., Ildefonso Cases, and Alfonso Valencia. 2004. “Gene Order in Prokaryotes: Conservation and Implications.” In *Molecules in Time and Space: Bacterial Shape, Division and Phylogeny*, edited by Miguel Vicente, Javier Tamames, Alfonso Valencia, and Jesús Mingorance, 209–37. Boston, MA: Springer US.

Gorle, Anil K., Amy L. Bottomley, Elizabeth J. Harry, J. Grant Collins, F. Richard Keene, and Clifford E. Woodward. 2017. “DNA Condensation in Live E. Coli Provides Evidence for Transertion.” *Molecular bioSystems* 13 (4): 677–80.

Grigoriev, Igor V., Henrik Nordberg, Igor Shabalov, Andrea Aerts, Mike Cantor, David Goodstein, Alan Kuo, et al. 2012. “The Genome Portal of the Department of Energy Joint Genome Institute.” *Nucleic Acids Research* 40 (Database issue): D26–32.

Gromadski, Kirill B., Hans-Joachim Wieden, and Marina V. Rodnina. 2002. “Kinetic Mechanism of Elongation Factor Ts-Catalyzed Nucleotide Exchange in Elongation Factor Tu.” *Biochemistry* 41 (1):

162–69.

Hahne, Hannes, Susanne Wolff, Michael Hecker, and Dörte Becher. 2008. "From Complementarity to Comprehensiveness--Targeting the Membrane Proteome of Growing Bacillus Subtilis by Divergent Approaches." *Proteomics* 8 (19): 4123–36.

Hamoen, Leendert W., Jean-Christophe Meile, Wouter de Jong, Philippe Noirot, and Jeff Errington. 2006. "SepF, a Novel FtsZ-Interacting Protein Required for a Late Step in Cell Division." *Molecular Microbiology* 59 (3): 989–99.

Helmann, John D. 2006. "Deciphering a Complex Genetic Regulatory Network: The Bacillus Subtilis sigmaW Protein and Intrinsic Resistance to Antimicrobial Compounds." *Science Progress* 89 (Pt 3-4): 243–66.

Hernández-Tamayo, Rogelio, Luis M. Oviedo-Bocanegra, Georg Fritz, and Peter L. Graumann. 2019. "Symmetric Activity of DNA Polymerases at and Recruitment of Exonuclease ExoR and of PolA to the Bacillus Subtilis Replication Forks." *Nucleic Acids Research*, June. https://doi.org/10.1093/nar/gkz554.

Hershberg, Ruth, and Dmitri A. Petrov. 2010. "Evidence That Mutation Is Universally Biased towards AT in Bacteria." *PLoS Genetics* 6 (9): e1001115.

Hess, Becky M., Junfeng Xue, Lye Meng Markillie, Ronald C. Taylor, H. Steven Wiley, Birgitte K. Ahring, and Bryan Linggi. 2013. "Coregulation of Terpenoid Pathway Genes and Prediction of Isoprene Production in Bacillus Subtilis Using Transcriptomics." *PloS One* 8 (6): e66104.

Hill, Norbert S., Paul J. Buske, Yue Shi, and Petra Anne Levin. 2013. "A Moonlighting Enzyme Links Escherichia Coli Cell Size with Central Metabolism." *PLoS Genetics* 9 (7): e1003663.

Hill, Norbert S., Ryosuke Kadoya, Dhruba K. Chattoraj, and Petra Anne Levin. 2012. "Cell Size and the Initiation of DNA Replication in Bacteria." *PLoS Genetics* 8 (3): e1002549.

Hobl, Birgit, and Matthias Mack. 2007. "The Regulator Protein PyrR of Bacillus Subtilis Specifically Interacts in Vivo with Three Untranslated Regions within Pyr mRNA of Pyrimidine Biosynthesis." *Microbiology* 153 (Pt 3): 693–700.

Hoffart, Eugenia, Sebastian Grenz, Julian Lange, Robert Nitschel, Felix Müller, Andreas Schwentner, André Feith, Mira Lenfers-Lücker, Ralf Takors, and Bastian Blombach. 2017. "High Substrate Uptake Rates Empower Vibrio Natriegens as Production Host for Industrial Biotechnology." *Applied and Environmental Microbiology* 83 (22). https://doi.org/10.1128/AEM.01614-17.

House, Christopher H., Matteo Pellegrini, and Sorel T. Fitz-Gibbon. 2014. "Genome-Wide Gene Order Distances Support Clustering the Gram-Positive Bacteria." *Frontiers in Microbiology* 5: 785.

Huang, Da Wei, Brad T. Sherman, and Richard A. Lempicki. 2009. "Systematic and Integrative Analysis of Large Gene Lists Using DAVID Bioinformatics Resources." *Nature Protocols* 4 (1): 44–57.

Huang, Xiaoluo, Daniela Pinto, Georg Fritz, and Thorsten Mascher. 2015. "Environmental Sensing in Actinobacteria: A Comprehensive Survey on the Signaling Capacity of This Phylum." *Journal of Bacteriology* 197 (15): 2517–35.

Huberts, Daphne H. E. W., and Ida J. van der Klei. 2010. "Moonlighting Proteins: An Intriguing Mode of Multitasking." *Biochimica et Biophysica Acta* 1803 (4): 520–25.

Huynen, M., B. Snel, W. Lathe 3rd, and P. Bork. 2000. "Predicting Protein Function by Genomic Context: Quantitative Evaluation and Qualitative Inferences." *Genome Research* 10 (8): 1204–10.

Itaya, M., A. Omori, S. Kanaya, R. J. Crouch, T. Tanaka, and K. Kondo. 1999. "Isolation of RNase H Genes That Are Essential for Growth of Bacillus Subtilis 168." *Journal of Bacteriology* 181 (7): 2118–23.

Itoh, T., K. Takemoto, H. Mori, and T. Gojobori. 1999. "Evolutionary Instability of Operon Structures Disclosed by Sequence Comparisons of Complete Microbial Genomes." *Molecular Biology and Evolution* 16 (3): 332–46.

Jacob, F., and J. Monod. 1961. "Genetic Regulatory Mechanisms in the Synthesis of Proteins." *Journal of Molecular Biology* 3 (June): 318–56.

Jain, Vikas, Manish Kumar, and Dipankar Chatterji. 2006. "ppGpp: Stringent Response and Survival." *Journal of Microbiology* 44 (1): 1–10.

Janga, Sarath Chandra, Warren F. Lamboy, Araceli M. Huerta, and Gabriel Moreno-Hagelsieb. 2006. “The Distinctive Signatures of Promoter Regions and Operon Junctions across Prokaryotes.” *Nucleic Acids Research* 34 (14): 3980–87.

Janosi, L., I. Shimizu, and A. Kaji. 1994. “Ribosome Recycling Factor (ribosome Releasing Factor) Is Essential for Bacterial Growth.” *Proceedings of the National Academy of Sciences of the United States of America* 91 (10): 4249–53.

Jogler, Christian, Jost Waldmann, Xiaoluo Huang, Mareike Jogler, Frank Oliver Glöckner, Thorsten Mascher, and Roberto Kolter. 2012. “Identification of Proteins Likely to Be Involved in Morphogenesis, Cell Division, and Signal Transduction in Planctomycetes by Comparative Genomics.” *Journal of Bacteriology* 194 (23): 6419–30.

Jordan, Sina, Anja Junker, John D. Helmann, and Thorsten Mascher. 2006. “Regulation of LiaRS-Dependent Gene Expression in Bacillus Subtilis: Identification of Inhibitor Proteins, Regulator Binding Sites, and Target Genes of a Conserved Cell Envelope Stress-Sensing Two-Component System.” *Journal of Bacteriology* 188 (14): 5153–66.

Joseph, Pascale, Gwennaele Fichant, Yves Quentin, and François Denizot. 2002. “Regulatory Relationship of Two-Component and ABC Transport Systems and Clustering of Their Genes in the Bacillus/Clostridium Group, Suggest a Functional Link between Them.” *Journal of Molecular Microbiology and Biotechnology* 4 (5): 503–13.

Kabeya, Yukihiro, Hiromitsu Nakanishi, Kenji Suzuki, Takanari Ichikawa, Youichi Kondou, Minami Matsui, and Shin-Ya Miyagishima. 2010. “The YlmG Protein Has a Conserved Function Related to the Distribution of Nucleoids in Chloroplasts and Cyanobacteria.” *BMC Plant Biology* 10 (April): 57.

Kaminishi, Tatsuya, Daniel N. Wilson, Chie Takemoto, Joerg M. Harms, Masahito Kawazoe, Frank Schluenzen, Kyoko Hanawa-Suetsugu, Mikako Shirouzu, Paola Fucini, and Shigeyuki Yokoyama. 2007. “A Snapshot of the 30S Ribosomal Subunit Capturing mRNA via the Shine-Dalgarno Interaction.” *Structure* 15 (3): 289–97.

Kapitonov, Vladimir V., and Jerzy Jurka. 2008. “A Universal Classification of Eukaryotic Transposable Elements Implemented in Repbase.” *Nature Reviews. Genetics*.

Keseler, Ingrid M., Amanda Mackie, Alberto Santos-Zavaleta, Richard Billington, César Bonavides-Martínez, Ron Caspi, Carol Fulcher, et al. 2017. “The EcoCyc Database: Reflecting New Knowledge about Escherichia Coli K-12.” *Nucleic Acids Research* 45 (D1): D543–50.

Kim, L., A. Mogk, and W. Schumann. 1996. “A Xylose-Inducible Bacillus Subtilis Integration Vector and Its Application.” *Gene* 181 (1-2): 71–76.

Kingston, A. W., C. Subramanian, and C. O. Rock. 2011. “A σW-dependent Stress Response in Bacillus Subtilis That Reduces Membrane Fluidity.” *Molecular*. https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-2958.2011.07679.x.

Kleerebezemab, M., P. Hols, and J. Hugenholtz. 2000. “Lactic Acid Bacteria as a Cell Factory: Rerouting of Carbon Metabolism in Lactococcus Lactis by Metabolic Engineering.” *Enzyme and Microbial Technology* 26 (9-10): 840–48.

Koch, A. L. 1985. “How Bacteria Grow and Divide in Spite of Internal Hydrostatic Pressure.” *Canadian Journal of Microbiology* 31 (12): 1071–84.

Koshland, Daniel E., Jr. 2002. “Special Essay. The Seven Pillars of Life.” *Science* 295 (5563): 2215–16.

Kriventseva, Evgenia V., Dmitry Kuznetsov, Fredrik Tegenfeldt, Mosè Manni, Renata Dias, Felipe A. Simão, and Evgeny M. Zdobnov. 2019. “OrthoDB v10: Sampling the Diversity of Animal, Plant, Fungal, Protist, Bacterial and Viral Genomes for Evolutionary and Functional Annotations of Orthologs.” *Nucleic Acids Research* 47 (D1): D807–11.

Kriventseva, Evgenia V., Fredrik Tegenfeldt, Tom J. Petty, Robert M. Waterhouse, Felipe A. Simão, Igor A. Pozdnyakov, Panagiotis Ioannidis, and Evgeny M. Zdobnov. 2015. “OrthoDB v8: Update of the Hierarchical Catalog of Orthologs and the Underlying Free Software.” *Nucleic Acids Research* 43 (Database issue): D250–56.

Kuroda, A., K. Nomura, R. Ohtomo, J. Kato, T. Ikeda, N. Takiguchi, H. Ohtake, and A. Kornberg. 2001. “Role of Inorganic Polyphosphate in Promoting Ribosomal Protein Degradation by the Lon Protease in E. Coli.” *Science* 293 (5530): 705–8.

Lachner, Monika, and Thomas Jenuwein. 2002. “The Many Faces of Histone Lysine Methylation.” *Current Opinion in Cell Biology* 14 (3): 286–98.

Laing, Emma, Vassilis Mersinias, Colin P. Smith, and Simon J. Hubbard. 2006. “Analysis of Gene Expression in Operons of Streptomyces Coelicolor.” *Genome Biology* 7 (6): R46.

Langmead, Ben, and Steven L. Salzberg. 2012. “Fast Gapped-Read Alignment with Bowtie 2.” *Nature Methods* 9 (4): 357–59.

Lathe, W. C., 3rd, B. Snel, and P. Bork. 2000. “Gene Context Conservation of a Higher Order than Operons.” *Trends in Biochemical Sciences* 25 (10): 474–79.

Lawrence, J. 1999. “Selfish Operons: The Evolutionary Impact of Gene Clustering in Prokaryotes and Eukaryotes.” *Current Opinion in Genetics & Development* 9 (6): 642–48.

Lawrence, J. G., and J. R. Roth. 1996. “Selfish Operons: Horizontal Transfer May Drive the Evolution of Gene Clusters.” *Genetics* 143 (4): 1843–60.

Lechner, Marcus, Sven Findeiss, Lydia Steiner, Manja Marz, Peter F. Stadler, and Sonja J. Prohaska. 2011. “Proteinortho: Detection of (co-)orthologs in Large-Scale Analysis.” *BMC Bioinformatics* 12 (April): 124.

Lewin, Gina R., Camila Carlos, Marc G. Chevrette, Heidi A. Horn, Bradon R. McDonald, Robert J. Stankey, Brian G. Fox, and Cameron R. Currie. 2016. “Evolution and Ecology of Actinobacteria and Their Bioenergy Applications.” *Annual Review of Microbiology* 70 (September): 235–54.

Lewis, P. J., S. D. Thaker, and J. Errington. 2000. “Compartmentalization of Transcription and Translation in Bacillus Subtilis.” *The EMBO Journal* 19 (4): 710–18.

Libby, Elizabeth A., Manuela Roggiani, and Mark Goulian. 2012. “Membrane Protein Expression Triggers Chromosomal Locus Repositioning in Bacteria.” *Proceedings of the National Academy of Sciences of the United States of America* 109 (19): 7445–50.

Li, Heng, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup. 2009. “The Sequence Alignment/Map Format and SAMtools.” *Bioinformatics* 25 (16): 2078–79.

Lin, Jie, and Ariel Amir. 2017. “The Effects of Stochasticity at the Single-Cell Level and Cell Size Control on the Population Growth.” *Cell Systems* 5 (4): 358–67.e4.

Lorenzo, V. de, and J. Pérez-Martín. 1996. “Regulatory Noise in Prokaryotic Promoters: How Bacteria Learn to Respond to Novel Environmental Signals.” *Molecular Microbiology* 19 (6): 1177–84.

Lu, Chung-Dar. 2006. “Pathways and Regulation of Bacterial Arginine Metabolism and Perspectives for Obtaining Arginine Overproducing Strains.” *Applied Microbiology and Biotechnology* 70 (3): 261–72.

Lu, M., J. L. Campbell, E. Boye, and N. Kleckner. 1994. “SeqA: A Negative Modulator of Replication Initiation in E. Coli.” *Cell* 77 (3): 413–26.

Luo, Hao, Yan Lin, Feng Gao, Chun-Ting Zhang, and Ren Zhang. 2014. “DEG 10, an Update of the Database of Essential Genes That Includes Both Protein-Coding Genes and Noncoding Genomic Elements.” *Nucleic Acids Research* 42 (Database issue): D574–80.

Maeda, H., N. Fujita, and A. Ishihama. 2000. “Competition among Seven Escherichia Coli Sigma Subunits: Relative Binding Affinities to the Core RNA Polymerase.” *Nucleic Acids Research* 28 (18): 3497–3503.

Mao, Fenglou, Phuongan Dam, Jacky Chou, Victor Olman, and Ying Xu. 2009. “DOOR: A Database for Prokaryotic Operons.” *Nucleic Acids Research* 37 (Database issue): D459–63.

Matsumoto, Kouji, Hiroshi Hara, Itzhak Fishov, Eugenia Mileykovskaya, and Vic Norris. 2015. “The Membrane: Transertion as an Organizing Principle in Membrane Heterogeneity.” *Frontiers in Microbiology* 6 (June): 572.

Ma, Zheng, Libin Tao, Andreas Bechthold, Xuping Shentu, Yalin Bian, and Xiaoping Yu. 2014. “Overexpression of Ribosome Recycling Factor Is Responsible for Improvement of Nucleotide Antibiotic-Toyocamycin in Streptomyces Diastatochromogenes 1628.” *Applied Microbiology and Biotechnology* 98 (11): 5051–58.

McCarn, D. F., R. A. Whitaker, J. Alam, J. M. Vrba, and S. E. Curtis. 1988. "Genes Encoding the Alpha, Gamma, Delta, and Four F0 Subunits of ATP Synthase Constitute an Operon in the Cyanobacterium Anabaena Sp. Strain PCC 7120." *Journal of Bacteriology* 170 (8): 3448–58.

McCLINTOCK, B. 1950. "The Origin and Behavior of Mutable Loci in Maize." *Proceedings of the National Academy of Sciences of the United States of America* 36 (6): 344–55.

McCormick, J. R., E. P. Su, A. Driks, and R. Losick. 1994. "Growth and Viability of Streptomyces Coelicolor Mutant for the Cell Division Gene ftsZ." *Molecular Microbiology* 14 (2): 243–54.

McIntyre, T. M., B. K. Chamberlain, R. E. Webster, and R. M. Bell. 1977. "Mutants of Escherichia Coli Defective in Membrane Phospholipid Synthesis. Effects of Cessation and Reinitiation of Phospholipid Synthesis on Macromolecular Synthesis and Phospholipid Turnover." *The Journal of Biological Chemistry* 252 (13): 4487–93.

McKenna, Aaron, Matthew Hanna, Eric Banks, Andrey Sivachenko, Kristian Cibulskis, Andrew Kernytsky, Kiran Garimella, et al. 2010. "The Genome Analysis Toolkit: A MapReduce Framework for Analyzing next-Generation DNA Sequencing Data." *Genome Research* 20 (9): 1297–1303.

Michalak, Pawel. 2008. "Coexpression, Coregulation, and Cofunctionality of Neighboring Genes in Eukaryotic Genomes." *Genomics* 91 (3): 243–48.

Michna, Raphael H., Bingyao Zhu, Ulrike Mäder, and Jörg Stülke. 2016. "SubtiWiki 2.0--an Integrated Database for the Model Organism Bacillus Subtilis." *Nucleic Acids Research* 44 (D1): D654–62.

Mingorance, Jesús, Javier Tamames, and Miguel Vicente. 2004. "Genomic Channeling in Bacterial Cell Division." *Journal of Molecular Recognition: JMR* 17 (5): 481–87.

Mitra, Anirban, Kandavelmani Angamuthu, Hanasoge Vasudevamurthy Jayashree, and Valakunja Nagaraja. 2009. "Occurrence, Divergence and Evolution of Intrinsic Terminators across Eubacteria." *Genomics* 94 (2): 110–16.

Mondal, Smarajit, Alexander V. Yakhnin, Aswathy Sebastian, Istvan Albert, and Paul Babitzke. 2016. "NusA-Dependent Transcription Termination Prevents Misregulation of Global Gene Expression." *Nature Microbiology* 1 (January): 15007.

Monds, Russell D., Timothy K. Lee, Alexandre Colavin, Tristan Ursell, Selwyn Quan, Tim F. Cooper, and Kerwyn Casey Huang. 2014. "Systematic Perturbation of Cytoskeletal Function Reveals a Linear Scaling Relationship between Cell Geometry and Fitness." *Cell Reports* 9 (4): 1528–37.

Moreno-Hagelsieb, Gabriel, and Julio Collado-Vides. 2002. "A Powerful Non-Homology Method for the Prediction of Operons in Prokaryotes." *Bioinformatics* 18 Suppl 1: S329–36.

Muro, Enrique M., Nancy Mah, Gabriel Moreno-Hagelsieb, and Miguel A. Andrade-Navarro. 2011. "The Pseudogenes of Mycobacterium Leprae Reveal the Functional Relevance of Gene Order within Operons." *Nucleic Acids Research* 39 (5): 1732–38.

Mushegian, A. R., and E. V. Koonin. 1996. "Gene Order Is Not Conserved in Bacterial Evolution." *Trends in Genetics: TIG* 12 (8): 289–90.

NCBI Resource Coordinators. 2018. "Database Resources of the National Center for Biotechnology Information." *Nucleic Acids Research* 46 (D1): D8–13.

Neef, Jolanda, Cristina Bongiorni, Vivianne J. Goosens, Brian Schmidt, and Jan Maarten van Dijl. 2017. "Intramembrane Protease RasP Boosts Protein Production in Bacillus." *Microbial Cell Factories* 16 (1): 57.

Nikolaichik, Y. A., and W. D. Donachie. 2000. "Conservation of Gene Order amongst Cell Wall and Cell Division Genes in Eubacteria, and Ribosomal Genes in Eubacteria and Eukaryotic Organelles." *Genetica* 108 (1): 1–7.

Nishibori, Ayako, Jin Kusaka, Hiroshi Hara, Masato Umeda, and Kouji Matsumoto. 2005. "Phosphatidylethanolamine Domains and Localization of Phospholipid Synthases in Bacillus Subtilis Membranes." *Journal of Bacteriology* 187 (6): 2163–74.

Noens, Elke E. E., and Juke S. Lolkema. 2017. "Convergent Evolution of the Arginine Deiminase Pathway: The ArcD and ArcE Arginine/ornithine Exchangers." *MicrobiologyOpen* 6 (1). https://doi.org/10.1002/mbo3.412.

Nomura, M., R. Gourse, and G. Baughman. 1984. "Regulation of the Synthesis of Ribosomes and

Ribosomal Components." *Annual Review of Biochemistry* 53: 75–117.

Noria, Stanislas, and Antoine Danchin. 2002. "Just so Genome Stories: What Does My Neighbor Tell Me?" *International Congress Series / Excerpta Medica*, December, 3–13.

Ohkubo, S., A. Muto, Y. Kawauchi, F. Yamao, and S. Osawa. 1987. "The Ribosomal Protein Gene Cluster of Mycoplasma Capricolum." *Molecular & General Genetics: MGG* 210 (2): 314–22.

Ohtani, N., M. Haruki, A. Muroya, M. Morikawa, and S. Kanaya. 2000. "Characterization of Ribonuclease HII from Escherichia Coli Overproduced in a Soluble Form." *Journal of Biochemistry* 127 (5): 895–99.

Oliveira, Pedro H., Marie Touchon, Jean Cury, and Eduardo P. C. Rocha. 2017. "The Chromosomal Organization of Horizontal Gene Transfer in Bacteria." *Nature Communications* 8 (1): 841.

Overbeek, R., M. Fonstein, M. D'Souza, G. D. Pusch, and N. Maltsev. 1999. "The Use of Gene Clusters to Infer Functional Coupling." *Proceedings of the National Academy of Sciences of the United States of America* 96 (6): 2896–2901.

Overbeek, Ross, Tadhg Begley, Ralph M. Butler, Jomuna V. Choudhuri, Han-Yu Chuang, Matthew Cohoon, Valérie de Crécy-Lagard, et al. 2005. "The Subsystems Approach to Genome Annotation and Its Use in the Project to Annotate 1000 Genomes." *Nucleic Acids Research* 33 (17): 5691–5702.

Ozaki, M., A. Wada, N. Fujita, and A. Ishihama. 1991. "Growth Phase-Dependent Modification of RNA Polymerase in Escherichia Coli." *Molecular & General Genetics: MGG* 230 (1-2): 17–23.

Paintdakhi, Ahmad, Bradley Parry, Manuel Campos, Irnov Irnov, Johan Elf, Ivan Surovtsev, and Christine Jacobs-Wagner. 2016. "Oufti: An Integrated Software Package for High-Accuracy, High-Throughput Quantitative Microscopy Analysis." *Molecular Microbiology* 99 (4): 767–77.

Pál, Csaba, and Laurence D. Hurst. 2004. "Evidence against the Selfish Operon Theory." *Trends in Genetics: TIG* 20 (6): 232–34.

Pamela, J. B, Brown, David, T, Kysela, Yves, V, and Brun. n.d. "Polarity and the Diversity of Growth Mechanisms in Bacteria | Elsevier Enhanced Reader." Accessed August 8, 2019. https://doi.org/10.1016/j.semcdb.2011.06.006.

Parrell, Daniel, Yang Zhang, Sandra Olenic, and Lee Kroos. 2017. "Bacillus Subtilis Intramembrane Protease RasP Activity in Escherichia Coli and In Vitro." *Journal of Bacteriology* 199 (19). https://doi.org/10.1128/JB.00381-17.

Parveen, Sadiya, and Manjula Reddy. 2017. "Identification of YfiH (PgeF) as a Factor Contributing to the Maintenance of Bacterial Peptidoglycan Composition." *Molecular Microbiology* 105 (5): 705–20.

Paul, Sandip, Samuel Million-Weaver, Sujay Chattopadhyay, Evgeni Sokurenko, and Houra Merrikh. 2013. "Accelerated Gene Evolution through Replication-Transcription Conflicts." *Nature* 495 (7442): 512–15.

Peters, Jason M., Alexandre Colavin, Handuo Shi, Tomasz L. Czarny, Matthew H. Larson, Spencer Wong, John S. Hawkins, et al. 2016. "A Comprehensive, CRISPR-Based Functional Analysis of Essential Genes in Bacteria." *Cell* 165 (6): 1493–1506.

Prescott, Elizabeth M., and Nick J. Proudfoot. 2002. "Transcriptional Collision between Convergent Genes in Budding Yeast." *Proceedings of the National Academy of Sciences of the United States of America* 99 (13): 8796–8801.

Price, Morgan N., Paramvir S. Dehal, and Adam P. Arkin. 2010. "FastTree 2--Approximately Maximum-Likelihood Trees for Large Alignments." *PloS One* 5 (3): e9490.

Price, Morgan N., Katherine H. Huang, Eric J. Alm, and Adam P. Arkin. 2005. "A Novel Method for Accurate Operon Predictions in All Sequenced Prokaryotes." *Nucleic Acids Research* 33 (3): 880–92.

Pucci, M. J., J. A. Thanassi, L. F. Discotto, R. E. Kessler, and T. J. Dougherty. 1997. "Identification and Characterization of Cell Wall-Cell Division Gene Clusters in Pathogenic Gram-Positive Cocci." *Journal of Bacteriology* 179 (17): 5632–35.

Qi, Ji, Bin Wang, and Bai-Iin Hao. 2004. "Whole Proteome Prokaryote Phylogeny without Sequence Alignment: A K-String Composition Approach." *Journal of Molecular Evolution* 58 (1): 1–11.

Quast, Christian, Elmar Pruesse, Pelin Yilmaz, Jan Gerken, Timmy Schweer, Pablo Yarza, Jörg Peplies, and Frank Oliver Glöckner. 2013. "The SILVA Ribosomal RNA Gene Database Project: Improved

Data Processing and Web-Based Tools." *Nucleic Acids Research* 41 (Database issue): D590–96.

Radeck, Jara, Korinna Kraft, Julia Bartels, Tamara Cikovic, Franziska Dürr, Jennifer Emenegger, Simon Kelterborn, et al. 2013. "The Bacillus BioBrick Box: Generation and Evaluation of Essential Genetic Building Blocks for Standardized Work with Bacillus Subtilis." *Journal of Biological Engineering* 7 (1): 29.

Ray, J. Christian J., and Oleg A. Igoshin. 2012. "Interplay of Gene Expression Noise and Ultrasensitive Dynamics Affects Bacterial Operon Organization." *PLoS Computational Biology* 8 (8): e1002672.

Real, Gonçalo, and Adriano O. Henriques. 2006. "Localization of the Bacillus Subtilis murB Gene within the Dcw Cluster Is Important for Growth and Sporulation." *Journal of Bacteriology* 188 (5): 1721–32.

Ream, David C., Asma R. Bankapur, and Iddo Friedberg. 2015. "An Event-Driven Approach for Studying Gene Block Evolution in Bacteria." *Bioinformatics* 31 (13): 2075–83.

Reiter, Lillian, Anne-Brit Kolstø, and Armin P. Piehler. 2011. "Reference Genes for Quantitative, Reverse-Transcription PCR in Bacillus Cereus Group Strains throughout the Bacterial Life Cycle." *Journal of Microbiological Methods* 86 (2): 210–17.

Rivas, E., R. J. Klein, T. A. Jones, and S. R. Eddy. 2001. "Computational Identification of Noncoding RNAs in E. Coli by Comparative Genomics." *Current Biology: CB* 11 (17): 1369–73.

Rocha, Eduardo P. C., and Antoine Danchin. 2002. "Base Composition Bias Might Result from Competition for Metabolic Resources." *Trends in Genetics: TIG* 18 (6): 291–94.

———. 2003. "Essentiality, Not Expressiveness, Drives Gene-Strand Bias in Bacteria." *Nature Genetics* 34 (4): 377–78.

Rogozin, Igor B., Kira S. Makarova, Janos Murvai, Eva Czabarka, Yuri I. Wolf, Roman L. Tatusov, Laszlo A. Szekely, and Eugene V. Koonin. 2002. "Connected Gene Neighborhoods in Prokaryotic Genomes." *Nucleic Acids Research* 30 (10): 2212–23.

Russell, J. B., and G. M. Cook. 1995. "Energetics of Bacterial Growth: Balance of Anabolic and Catabolic Reactions." *Microbiological Reviews* 59 (1): 48–62.

Salgado, H., G. Moreno-Hagelsieb, T. F. Smith, and J. Collado-Vides. 2000. "Operons in Escherichia Coli: Genomic Analyses and Predictions." *Proceedings of the National Academy of Sciences of the United States of America* 97 (12): 6652–57.

Sander, Timur, Chun Ying Wang, Timo Glatter, and Hannes Link. 2019. "CRISPRi-Based Downregulation of Transcriptional Feedback Improves Growth and Metabolism of Arginine Overproducing E. Coli." *ACS Synthetic Biology* 8 (9): 1983–90.

Santos-Beneit, Fernando, Jing-Ying Gu, Ulrich Stimming, and Jeff Errington. 2017. "ylmD and ylmE Genes Are Dispensable for Growth, Cross-Wall Formation and Sporulation in Streptomyces Venezuelae." *Heliyon* 3 (11): e00459.

Santos-Rosa, Helena, Robert Schneider, Andrew J. Bannister, Julia Sherriff, Bradley E. Bernstein, N. C. Tolga Emre, Stuart L. Schreiber, Jane Mellor, and Tony Kouzarides. 2002. "Active Genes Are Tri-Methylated at K4 of Histone H3." *Nature* 419 (6905): 407–11.

Sargent, M. G. 1975. "Control of Cell Length in Bacillus Subtilis." *Journal of Bacteriology* 123 (1): 7–19.

Sauer, Christopher, Simon Syvertsson, Laura C. Bohorquez, Rita Cruz, Colin R. Harwood, Tjeerd van Rij, and Leendert W. Hamoen. 2016. "Effect of Genome Position on Heterologous Gene Expression in Bacillus Subtilis: An Unbiased Analysis." *ACS Synthetic Biology* 5 (9): 942–47.

Schaechter, M., O. Maaloe, and N. O. Kjeldgaard. 1958. "Dependency on Medium and Temperature of Cell Size and Chemical Composition during Balanced Grown of Salmonella Typhimurium." *Journal of General Microbiology* 19 (3): 592–606.

Schmidt, M. C., and M. J. Chamberlin. 1987. "nusA Protein of Escherichia Coli Is an Efficient Transcription Termination Factor for Certain Terminator Sites." *Journal of Molecular Biology* 195 (4): 809–18.

Schneider, B. L., A. K. Kiupakis, and L. J. Reitzer. 1998. "Arginine Catabolism and the Arginine Succinyltransferase Pathway in Escherichia Coli." *Journal of Bacteriology* 180 (16): 4278–86.

Scott, Matthew, and Terence Hwa. 2011. "Bacterial Growth Laws and Their Applications." *Current*

Opinion in Biotechnology 22 (4): 559–65.

Sievers, Fabian, Andreas Wilm, David Dineen, Toby J. Gibson, Kevin Karplus, Weizhong Li, Rodrigo Lopez, et al. 2011. “Fast, Scalable Generation of High-Quality Protein Multiple Sequence Alignments Using Clustal Omega.” *Molecular Systems Biology* 7 (October): 539.

Si, Fangwei, Dongyang Li, Sarah E. Cox, John T. Sauls, Omid Azizi, Cindy Sou, Amy B. Schwartz, et al. 2017. “Invariance of Initiation Mass and Predictability of Cell Size in Escherichia Coli.” *Current Biology: CB* 27 (9): 1278–87.

Silva, Paula Renata Alves da, Marcia Soares Vidal, Cleiton de Paula Soares, Valéria Polese, Jean Luís Simões-Araújo, and José Ivo Baldani. 2016. “Selection and Evaluation of Reference Genes for RT-qPCR Expression Studies on Burkholderia Tropica Strain Ppe8, a Sugarcane-Associated Diazotrophic Bacterium Grown with Different Carbon Sources or Sugarcane Juice.” *Antonie van Leeuwenhoek* 109 (11): 1493–1502.

Simon, Meinhard, and Farooq Azam. 1989. “Protein Content and Protein Synthesis Rates of Planktonic Marine Bacteria.” *Marine Ecology Progress Series. Oldendorf* 51 (3): 201–13.

Siow, Cheuk C., Sian R. Nieduszynska, Carolin A. Müller, and Conrad A. Nieduszynski. 2012. “OriDB, the DNA Replication Origin Database Updated and Extended.” *Nucleic Acids Research* 40 (Database issue): D682–86.

Sirand-Pugnet, Pascal, Christine Citti, Aurélien Barré, and Alain Blanchard. 2007. “Evolution of Mollicutes: Down a Bumpy Road with Twists and Turns.” *Research in Microbiology* 158 (10): 754–66.

Sivy, Tami L., Megan C. Shirk, and Ray Fall. 2002. “Isoprene Synthase Activity Parallels Fluctuations of Isoprene Release during Growth of Bacillus Subtilis.” *Biochemical and Biophysical Research Communications* 294 (1): 71–75.

Skarstad, Kirsten, and Tsutomu Katayama. 2013. “Regulating DNA Replication in Bacteria.” *Cold Spring Harbor Perspectives in Biology* 5 (4): a012922.

Smith, Gerald R. 2012. “How RecBCD Enzyme and Chi Promote DNA Break Repair and Recombination: A Molecular Biologist’s View.” *Microbiology and Molecular Biology Reviews: MMBR* 76 (2): 217–28.

Snel, B., P. Bork, and M. A. Huynen. 1999. “Genome Phylogeny Based on Gene Content.” *Nature Genetics* 21 (1): 108–10.

Snel, B., G. Lehmann, P. Bork, and M. A. Huynen. 2000. “STRING: A Web-Server to Retrieve and Display the Repeatedly Occurring Neighbourhood of a Gene.” *Nucleic Acids Research* 28 (18): 3442–44.

Soler-Bistué, Alfonso, Michaël Timmermans, and Didier Mazel. 2017. “The Proximity of Ribosomal Protein Genes to oriC Enhances Vibrio Cholerae Fitness in the Absence of Multifork Replication.” *mBio* 8 (1). https://doi.org/10.1128/mBio.00097-17.

Solomon, J. M., and A. D. Grossman. 1996. “Who’s Competent and When: Regulation of Natural Genetic Competence in Bacteria.” *Trends in Genetics: TIG* 12 (4): 150–55.

Stalon, V., C. Vander Wauven, P. Momin, and C. Legrain. 1987. “Catabolism of Arginine, Citrulline and Ornithine by Pseudomonas and Related Bacteria.” *Journal of General Microbiology* 133 (9): 2487–95.

Staroń, Anna, Heidi J. Sofia, Sascha Dietrich, Luke E. Ulrich, Heiko Liesegang, and Thorsten Mascher. 2009. “The Third Pillar of Bacterial Signal Transduction: Classification of the Extracytoplasmic Function (ECF) Sigma Factor Protein Family.” *Molecular Microbiology* 74 (3): 557–81.

Stuessy, Tod F., and Christiane König. 2008. “Patrocladistic Classification.” *Taxon* 57 (2): 594–601.

Sukhodolets, Maxim V., and Susan Garges. 2003. “Interaction of Escherichia Coli RNA Polymerase with the Ribosomal Protein S1 and the Sm-like ATPase Hfq.” *Biochemistry* 42 (26): 8022–34.

Sun, Jianteng, Lili Pan, and Lizhong Zhu. 2018. “Formation of Hydroxylated and Methoxylated Polychlorinated Biphenyls by Bacillus Subtilis: New Insights into Microbial Metabolism.” *The Science of the Total Environment* 613-614 (February): 54–61.

Suyama, M., and P. Bork. 2001. “Evolution of Prokaryotic Gene Order: Genome Rearrangements in

Closely Related Species." *Trends in Genetics: TIG* 17 (1): 10–13.

Szadkowski, Dobromir, Andrea Harms, Luis António Menezes Carreira, Manon Wigbers, Anna Potapova, Kristin Wuichet, Daniela Keilberg, Ulrich Gerland, and Lotte Søgaard-Andersen. 2019. "Spatial Control of the GTPase MglA by Localized RomR-RomX GEF and MglB GAP Activities Enables Myxococcus Xanthus Motility." *Nature Microbiology* 4 (8): 1344–55.

Szklarczyk, Damian, Annika L. Gable, David Lyon, Alexander Junge, Stefan Wyder, Jaime Huerta-Cepas, Milan Simonovic, et al. 2019. "STRING v11: Protein-Protein Association Networks with Increased Coverage, Supporting Functional Discovery in Genome-Wide Experimental Datasets." *Nucleic Acids Research* 47 (D1): D607–13.

Taheri-Araghi, Sattar, Serena Bradde, John T. Sauls, Norbert S. Hill, Petra A. Levin, Johan Paulsson, Massimo Vergassola, and Suckjoon Jun. 2015. "Cell-Size Control and Homeostasis in Bacteria." *Current Biology: CB* 25 (3): 385–91.

Tamames, J., M. González-Moreno, J. Mingorance, A. Valencia, and M. Vicente. 2001. "Bringing Gene Order into Bacterial Shape." *Trends in Genetics: TIG* 17 (3): 124–26.

Tanimoto, K., Q. Liu, J. Bungert, and J. D. Engel. 1999. "Effects of Altered Gene Order or Orientation of the Locus Control Region on Human Beta-Globin Gene Expression in Mice." *Nature* 398 (6725): 344–48.

Tatusov, R. L., M. Y. Galperin, D. A. Natale, and E. V. Koonin. 2000. "The COG Database: A Tool for Genome-Scale Analysis of Protein Functions and Evolution." *Nucleic Acids Research* 28 (1): 33–36.

The Gene Ontology Consortium. 2019. "The Gene Ontology Resource: 20 Years and Still GOing Strong." *Nucleic Acids Research* 47 (D1): D330–38.

Thorvaldsdóttir, Helga, James T. Robinson, and Jill P. Mesirov. 2013. "Integrative Genomics Viewer (IGV): High-Performance Genomics Data Visualization and Exploration." *Briefings in Bioinformatics* 14 (2): 178–92.

Tian, G., D. Lim, J. D. Oppenheim, and W. K. Maas. 1994. "Explanation for Different Types of Regulation of Arginine Biosynthesis in Escherichia Coli B and Escherichia Coli K12 Caused by a Difference between Their Arginine Repressors." *Journal of Molecular Biology* 235 (1): 221–30.

Tillier, E. R., and R. A. Collins. 2000. "Genome Rearrangement by Replication-Directed Translocation." *Nature Genetics* 26 (2): 195–97.

Toledo-Arana, Alejandro, Francis Repoila, and Pascale Cossart. 2007. "Small Noncoding RNAs Controlling Pathogenesis." *Current Opinion in Microbiology* 10 (2): 182–88.

Tosa, T., and L. I. Pizer. 1971. "Effect of Serine Hydroxamate on the Growth of Escherichia Coli." *Journal of Bacteriology* 106 (3): 966–71.

Touchon, Marie, and Eduardo P. C. Rocha. 2016. "Coevolution of the Organization and Structure of Prokaryotic Genomes." *Cold Spring Harbor Perspectives in Biology* 8 (1): a018168.

Turner, R. J., Y. Lu, and R. L. Switzer. 1994. "Regulation of the Bacillus Subtilis Pyrimidine Biosynthetic (pyr) Gene Cluster by an Autogenous Transcriptional Attenuation Mechanism." *Journal of Bacteriology* 176 (12): 3708–22.

Vicente, M., M. J. Gomez, and J. A. Ayala. 1998. "Regulation of Transcription of Cell Division Genes in the Escherichia Coli Dcw Cluster." *Cellular and Molecular Life Sciences: CMLS* 54 (4): 317–24.

Vieira-Silva, Sara, and Eduardo P. C. Rocha. 2010. "The Systemic Imprint of Growth and Its Uses in Ecological (meta)genomics." *PLoS Genetics* 6 (1): e1000808.

Wallden, Mats, David Fange, Ebba Gregorsson Lundius, Özden Baltekin, and Johan Elf. 2016. "The Synchronization of Replication and Division Cycles in Individual E. Coli Cells." *Cell* 166 (3): 729–39.

Wassarman, K. M., F. Repoila, C. Rosenow, G. Storz, and S. Gottesman. 2001. "Identification of Novel Small RNAs Using Comparative Genomics and Microarrays." *Genes & Development* 15 (13): 1637–51.

Weart, Richard B., Amy H. Lee, An-Chun Chien, Daniel P. Haeusser, Norbert S. Hill, and Petra Anne Levin. 2007. "A Metabolic Sensor Governing Cell Size in Bacteria." *Cell* 130 (2): 335–47.

Weber, Ernst, Carola Engler, Ramona Gruetzner, Stefan Werner, and Sylvestre Marillonnet. 2011. "A

Modular Cloning System for Standardized Assembly of Multigene Constructs." *PloS One* 6 (2): e16765.

Wells, Jonathan N., L. Therese Bergendahl, and Joseph A. Marsh. 2016. "Operon Gene Order Is Optimized for Ordered Protein Complex Assembly." *Cell Reports* 14 (4): 679–85.

Westfall, Corey S., and Petra Anne Levin. 2017. "Bacterial Cell Size: Multifactorial and Multifaceted." *Annual Review of Microbiology* 71 (September): 499–517.

West, Stephen C. 2003. "Molecular Views of Recombination Proteins and Their Control." *Nature Reviews. Molecular Cell Biology* 4 (6): 435–45.

Wheeler, R. T., and L. Shapiro. 1997. "Bacterial Chromosome Segregation: Is There a Mitotic Apparatus?" *Cell* 88 (5): 577–79.

Wilkins, Joanna C., David Beighton, and Karen A. Homer. 2003. "Effect of Acidic pH on Expression of Surface-Associated Proteins of Streptococcus Oralis." *Applied and Environmental Microbiology* 69 (9): 5290–96.

Wold, Sture, Kirsten Skarstad, Harald B. Steen, Trond Stokke, and Erik Boye. 1994. "The Initiation Mass for DNA Replication in Escherichia Coli K-12 Is Dependent on Growth Rate." *The EMBO Journal* 13 (9): 2097–2102.

Wolf, Y. I., I. B. Rogozin, N. V. Grishin, R. L. Tatusov, and E. V. Koonin. 2001. "Genome Trees Constructed Using Five Different Approaches Suggest New Major Bacterial Clades." *BMC Evolutionary Biology* 1 (October): 8.

Wolf, Y. I., I. B. Rogozin, A. S. Kondrashov, and E. V. Koonin. 2001. "Genome Alignment, Evolution of Prokaryotic Genome Organization, and Prediction of Gene Function Using Genomic Context." *Genome Research* 11 (3): 356–72.

Xue, Junfeng, and Birgitte K. Ahring. 2011. "Enhancing Isoprene Production by Genetic Modification of the 1-Deoxy-D-Xylulose-5-Phosphate Pathway in Bacillus Subtilis." *Applied and Environmental Microbiology* 77 (7): 2399–2405.

Yusupova, Gulnara, Lasse Jenner, Bernard Rees, Dino Moras, and Marat Yusupov. 2006. "Structural Basis for Messenger RNA Movement on the Ribosome." *Nature* 444 (7117): 391–94.

Zegerman, Philip, Benito Canas, Darryl Pappin, and Tony Kouzarides. 2002. "Histone H3 Lysine 4 Methylation Disrupts Binding of Nucleosome Remodeling and Deacetylase (NuRD) Repressor Complex." *The Journal of Biological Chemistry* 277 (14): 11621–24.

Zellmeier, Stephan, Claudia Hofmann, Sylvia Thomas, Thomas Wiegert, and Wolfgang Schumann. 2005. "Identification of sigma(V)-Dependent Genes of Bacillus Subtilis." *FEMS Microbiology Letters* 253 (2): 221–29.

Zhang, Shumeng, Xinfeng Li, Xun Wang, Zhou Li, and Jin He. 2016. "The Two-Component Signal Transduction System YvcPQ Regulates the Bacterial Resistance to Bacitracin in Bacillus Thuringiensis." *Archives of Microbiology* 198 (8): 773–84.

Zheng, Xiao-Yu, and Erin K. O'Shea. 2017. "Cyanobacteria Maintain Constant Protein Concentration despite Genome Copy-Number Variation." *Cell Reports* 19 (3): 497–504.

Zheng, Yu, Brian P. Anton, Richard J. Roberts, and Simon Kasif. 2005. "Phylogenetic Detection of Conserved Gene Clusters in Microbial Genomes." *BMC Bioinformatics* 6 (October): 243.

Zhu, Bingyao, and Jörg Stülke. 2018. "SubtiWiki in 2018: From Genes and Proteins to Functional Network Annotation of the Model Organism Bacillus Subtilis." *Nucleic Acids Research* 46 (D1): D743–48.

Zumkeller, Celine, Daniel Schindler, and Torsten Waldminghaus. 2018. "Modular Assembly of Synthetic Secondary Chromosomes." In *Bacterial Chromatin: Methods and Protocols*, edited by Remus T. Dame, 71–94. New York, NY: Springer New York.

Zweers, Jessica C., Pierre Nicolas, Thomas Wiegert, Jan Maarten van Dijl, and Emma L. Denham. 2012. "Definition of the σ(W) Regulon of Bacillus Subtilis in the Absence of Stress." *PloS One* 7 (11): e48471.

8. Appendix

8.1 Bacterial strains, plasmids and primers used in this study

B.subtilis strains

ID	Genotype	Resistence	Comment
GFB0058	W168 uppsUTR::cat	Cm^R	Readthrough Blocked
GFB0097	W168 amyE::PDG380+pXylA+uppS->rasP	Cm^R	Operon Analysis
GFB0113	W168 amyE::PDG380+pXylA+uppS->rasP upps->rasp::cat PxylA-proS	Cm^R	Split Cluster
GFB0057	W168 lacA::dCas9 erm	Erm^R	Operon Analysis
GFB0074	W168 lacA::dCas9 erm AmyE::(rpsB-sgRNA cat)	Erm^R/Cm^R	Operon Analysis
GFB0075	W168 lacA::dCas9 erm AmyE::(tsf-sgRNA cat)	Erm^R/Cm^R	Operon Analysis
GFB0076	W168 lacA::dCas9 erm AmyE::(pyrH-sgRNA cat)	Erm^R/Cm^R	Operon Analysis
GFB0077	W168 lacA::dCas9 AmyE::(frr sgRNA cat)	Erm^R/Cm^R	Operon Analysis
GFB0078	W168 lacA::dCas9 AmyE::(uppS sgRNA cat)	Erm^R/Cm^R	Operon Analysis
GFB0079	W168 lacA::dCas9 AmyE::(cdsA sgRNA cat)	Erm^R/Cm^R	Operon Analysis
GFB0080	W168 lacA::dCas9 AmyE::(dxr sgRNA cat)	Erm^R/Cm^R	Operon Analysis
GFB0081	W168 lacA::dCas9 AmyE::(rasP sgRNA cat)	Erm^R/Cm^R	Operon Analysis
GFB0101	W168 polC-(tet array cat PxylA-tetR-YFP)	Erm^R/Cm^R	Transertion
GFB0104	W168 sacA:::cat-Pupps20-lux	Cm^R	Pupps Promoter Fusion
GFB0105	W168 sacA:::cat-Pupps80-lux	Cm^R	Pupps Promoter Fusion
GFB0106	W168 sacA:::cat-Pupps180-lux	Cm^R	Pupps Promoter Fusion
GFB0107	W168 sacA:::cat-Pupps120-lux	Cm^R	Pupps Promoter Fusion
GFB0108	W168 sacA:::cat-Pupps150-lux	Cm^R	Pupps Promoter Fusion
GFB0110	W168 sacA:::cat-Pupps40-lux	Cm^R	Pupps Promoter Fusion
GFB0111	W168 sacA:::cat-Pupps100-lux	Cm^R	Pupps Promoter Fusion
GFB0112	W168 sacA:::cat-Pupps200-lux	Cm^R	Pupps Promoter Fusion
GFB0115	W168 sacA:::cat-PuppS(180-140)-lux	Cm^R	Pupps Promoter Fusion
GFB0123	W168 sacA:::cat-PuppS(180-140)-Pupps20-lux	Cm^R	Pupps Promoter Fusion
GFB0124	W168 sacA:::cat-PuppS(180-140)-Pupps40-lux	Cm^R	Pupps Promoter Fusion
GFB0125	W168 sacA:::cat-PuppS(180-140)-Pupps80-lux	Cm^R	Pupps Promoter Fusion
GFB0126	W168 sacA:::cat-PuppS(180-140)-Pupps100-lux	Cm^R	Pupps Promoter Fusion
GFB0127	W168 sacA:::cat-PuppS(180-140)-Pupps120-lux	Cm^R	Pupps Promoter Fusion
GFB0128	W168 sacA:::cat-PuppS(180-140)-Pupps140-lux	Cm^R	Pupps Promoter Fusion
GFB0129	W168 sacA:::cat-PuppS(180-140)-Pupps180-lux	Cm^R	Pupps Promoter Fusion
GFB0130	W168 ΔrasP::spec-Pxyla	Spc^R	RasP knock out
GFB0153	W168 ΔrasP::spec-Pxyla sac::(Pliag-rasP erm)	Erm^R	RasP complement

7 - Table 8.1 – B. subtilis strains utilized in this study

Appendix

The name, the genotype, the antibiotic resistance and a description of the strain are reported. The strains are listed in alphabetical order. Cm^R: chloramphenicol resistance, Erm^R: erythromycin resistance, Spc^R: spectinomycin resistance.

Level 0 library

Level 0 ID	Name	Genetic Part	Vector	Donor	Primer forward	Primer reverse	Reference
8	pSV0-9_003	RBS strong (st8) corrected	pICH41246	Synthetic DNA	-	-	Robert Luis Vellanoweth and Jesse C. Rabinowitz
16	pSV0-11_001	L3S2P21 used in P...lux/gfp	pICH41276	Synthetic DNA	-	-	Ying-Ja Chen et al.
56	pSV0-11_006	dummy terminator	pICH41276	Synthetic DNA	-	-	AG Fritz, unpublished
74	pSV0-14_003	du15	pICH41295	Synthetic DNA	-	-	AG Fritz, unpublished
75	pSV0-15_022	du15	pICH41308	Synthetic DNA	-	-	AG Fritz, unpublished
82	pSV0-15_029	As20_992	pICH41308	Pseudomonas fluorescens Pf-5	-	-	Virgil A Rhodius et al.
100	pAT0-14_004	sacA front homologous region	pICH41295	B. subtilis	-	-	Radeck et al. 2013 pBs3C
101	pAT0-15_035	cat (chloramphenicol resistance) on minus strand	pICH41308	pBsC3lux	-	-	Radeck et al. 2013 pBs3C
103	pAT0-9_004	RBS (Ribosome binding site)	pICH41246	Synthetic DNA	-	-	Radeck et al. in front of lux operon in pBs3C
104	pAT0-15_036	lux operon (Photorhabdus luminescence)	pICH41308	Photorhabdus luminescens	-	-	Radeck et al. 2013 pBs3C
105	pAT0-11_016	sacA back homologous region	pICH41276	B. subtilis	-	-	Radeck et al. 2013 pBs3C
107	plAS0-14_005	LacA Left homologous region	pICH41295	B. subtilis	-	-	AG Fritz, unpublished
115	pAT0-15_037	spc	pICH41308	pBs4S	-	-	Radeck et al. 2013 pBS4S
116	pAT0-15_038	erm	pICH41308	pBs2E	-	-	Radeck et al. 2013 pBS2E
119	pJM0-9_005	du15	plCH41246	Synthetic DNA	-	-	AG Fritz, unpublished
122	pJM0-11_019	LacA Right homologous region	pICH41276	B. subtilis	-	-	AG Fritz, unpublished
128	pAS0-11-020	Upps 5' HR	pICH41276	B. subtilis	GF0382, GF0384	GF0383, GF0385	This study
130	pAS0-1-035	UppsP	pICH41233	B. subtilis	-	-	This study

178	pJM0-1_038	PliaG	plCH41233	B. subtilis	-	-	AG Fritz, unpublished
180	pJM0-1_040	PxylA	plCH41233	B. subtilis	-	-	AG Fritz, unpublished
186	pAS0-14_13	frr HR longer	pICH41295	B. subtilis	GF0386, GF0388, GF0390	GF0387, GF0389, GF0391	This study
201	pAS0-14_14	Post PolC homology region left	pICH41295	B. subtilis	GF0607	GF0608	This study
205	pAS0-15-55	TetYFP (from plau53)	pICH41308	plau53	GF0688	GF0689	This study
207	pAS0-15-56	tet array (from plau43)	pICH41308	plau43	GF0617	GF0618	This study
208	pAS0-11_024	Post PolC homology region right	pICH41276	B. subtilis	GF0609	GF0610	This study
219	pAS0-11_026	rasP HR R	pICH41276	B. subtilis	GF0702	GF0703	This study
220	pAS0-14_17	20bp UppS Promotor + RBS	pICH41295	B. subtilis	GF0706	GF0707	This study
221	pAS0-14_18	40bp UppS Promotor + RBS	pICH41295	B. subtilis	GF0708	GF0709	This study
223	pAS0-14_20	80bp UppS Promotor + RBS	pICH41295	B. subtilis	GF0712	GF0705	This study
224	pAS0-14_21	100bp UppS Promotor + RBS	pICH41295	B. subtilis	GF0713	GF0705	This study
225	pAS0-14_22	120bp UppS Promotor + RBS	pICH41295	B. subtilis	GF0714	GF0705	This study
226	pAS0-14_23	140bp UppS Promotor + RBS	pICH41295	B. subtilis	GF0715	GF0705	This study
228	pAS0-14_25	180bp UppS Promotor + RBS	pICH41295	B. subtilis	GF0717	GF0705	This study
229	pAS0-14_26	200bp UppS Promotor + RBS	pICH41295	B. subtilis	GF0718	GF0705	This study
261	pJB0-14_34	rasP HR Left	pICH41295	B. subtilis	GF0759, GF0761	GF0760, GF0762	This study
262	pAN0-1_55	Pupps no UTR	plCH41233	B. subtilis	GF0765	GF0766	This study
293	pJB0-15_061	rasP Gene	pICH41308	B. subtilis	GF0850, GF0843	GF0842, GF0844	This study

8 - Table 8.2 - Level 0 part library

MoClo-encoded level 0 parts either generated or used in this study. The parts of listed in numerical order. Internal name abbreviations: (pAS) generated by Andre Sim, (pSV) generated by Stefano Vecchione, (pAT) generated by Anika Thorhauer, (pJM) generated by Julia Manning, (pJB) generated by Jessica Bzdok, (pAN) generated by Annis Newman. The original name of MocClo destination vectors, in which the parts are encoded as well as the original names of the donor plasmids and primers used for PCR-amplification, or oligonucleotide annealing are indicated.

Level 1 library				
Level 1 ID	Name	Vector	Level 1 Doner parts	Description
149	pJM1-1L_0046	plCH47732	100 + 101 + 16	sacA integration R
189	pJM1-4L_0006	plCH47761	74 + 75 + 105	sacA integration L
196	pAS1-2L-0030	plCH47742	180 + 119 + 75 + 56	PxylA promoter
198	pAS1-3L-0064	plCH47751	130 + 119 + 75 + 128	UTR uppS integration
266	pAS1-1R-0004	plCH47802	74 + 82 + 56	cat
300	pAS-1-6L-0005	plCH47861	186 + 75 + 56	UTR uppS integration
332	pAS1-1L_0071	plCH47732	201 + 207 + 56	Transertion 1
333	pAS1-3L-0069	plCH47751	180 + 103 + 205 + 16	Transertion 3
334	pAS1-4L_0009	plCH47761	74 + 75 + 208	Transertion 4
368	pAS1-2L_0071	plCH47742	74 + 101 + 16	Transertion 2
382	pJB1-1L_0091	plCH47732	107 + 75 + 56	lacA integration L
384	pAS1-1L_0091	plCH47732	186 + 101 + 56	UTR uppS integration
385	pAS1-2L_0073	plCH47742	180 + 119 + 75 + 261	proS expresssion control
386	pAS1-1L_0093	plCH47732	100 + 101 + 16	sacA integration L with cat
399	pAS1-2L_0075	plCH47742	220 + 104 + 16	20bp UppS Promotor + RBS
400	pAS1-2L_0076	plCH47742	221 + 104 + 16	40bp UppS Promotor + RBS
402	pAS1-2L_0078	plCH47742	223 + 104 + 16	80bp UppS Promotor + RBS
403	pAS1-2L_0079	plCH47742	224 + 104 + 16	100bp UppS Promotor + RBS
404	pAS1-2L_0080	plCH47742	225 + 104 + 16	120bp UppS Promotor + RBS
405	pAS1-2L_0081	plCH47742	226 + 104 + 16	140bp UppS Promotor + RBS
407	pAS1-2L_0083	plCH47742	228 + 104 + 16	180bp UppS Promotor + RBS
440	pJB1-4L_0014	plCH47761	74 + 116 + 122	lacA integration R
444	pAN1-2L_0105	plCH47742	262 + 103 + 104 + 16	Pupps no UTR test
445	pJB1_1L_0111	plCH47732	261 + 115 + 16	for rasP knockout
446	pJB1-2L_0106	plCH47742	74 + 75 +219	for rasP knockout
454	pAN1-3L_0074	plCH47751	226 + 104 + 16	140bp PuppS Promoter Fusion
455	pAN1-3L_0075	plCH47751	228 + 104 + 16	180bp Pupps Promoter Fusion
456	pAN1-3L_0076	plCH47751	220 + 104 + 16	20bp Pupps Promoter Fusion
457	pAN1-3L_0077	plCH47751	221 + 104 + 16	40bp Pupps Promoter Fusion
458	pAN1-3L_0078	plCH47751	223 + 104 + 16	80bp Pupps Promoter Fusion
459	pAN1-3L_0079	plCH47751	224 + 104 + 16	100bp Pupps Promoter Fusion
460	pAN1-3L_0080	plCH47751	225 + 104 + 16	120bp Pupps Promoter Fusion
514	pJB1-2L_0113	plCH47742	178 + 119 + 75 + 56	PliaG promoter
515	pJB1-3L_0086	plCH47751	96 + 8 + 293 + 16	rasP

9 - Table 8.3 - Level 1 part library

MoClo-encoded level 1 parts either generated or used in this study. The parts of listed in numerical order. Internal name abbreviations: (pAS) generated by Andre Sim, (pJM) generated by Julia Manning, (pJB) generated by Jessica Bzdok, (pAN) generated by Annis Newman. The original name of MocClo destination vectors, in which the parts are encoded as well as the original names of the donor plasmids and primers used for PCR-amplification, or oligonucleotide annealing are indicated.

Level M library				
Level M ID	Name	Vector	Level 1 Doner parts	Description
240	**pASM-6_008**	pAGM8081	384 + 196 + 198	Readthrough Block
298	**pASM-1_062**	pAGM8031	332 + 368 + 333 + 334	Transertion
354	**pASM-1_92**	pAGM8031	386 + 402+ 189	PuppS 80
355	**pASM-1_93**	pAGM8031	386 + 404+ 189	PuppS 120
356	**pASM-1_94**	pAGM8031	386 + 405 + 189	PuppS 140
357	**pASM-1_95**	pAGM8031	386 + 407 + 189	PuppS 180
358	**pASM-1_96**	pAGM8031	386 + 399 + 189	PuppS 20
393	**pASM-1_98**	pAGM8031	386 +400 + 189	PuppS 40
394	**pASM-1_99**	pAGM8031	386 + 403 + 189	PuppS 100
399	**pASM-1_104**	pAGM8031	386 + 413 + 189	PuppS 200
401	**pASM-1_106**	pAGM8031	384 + 385	uppS-rasP deletion construct
412	**pANM-1_115**	pAGM8031	149 + 461 + 120	Pupps no UTR test
413	**pJBM-1_116**	pAGM8031	445 + 446	rasP knock out
418	**pANM-1_121**	pAGM8031	149 + 461 + 460 + 189	120bp PuppS Promoter Fusion
419	**pANM-1_122**	pAGM8031	149 + 461 + 457 + 189	40bp Pupps Promoter Fusion
420	**pANM-1_123**	pAGM8031	149 + 461 + 458 + 189	80bp Pupps Promoter Fusion
421	**pANM-1_124**	pAGM8031	149 + 461 + 454+ 189	140bp Pupps Promoter Fusion
422	**pANM-1_125**	pAGM8031	149 + 461 + 455+ 189	180bp Pupps Promoter Fusion
423	**pANM-1_126**	pAGM8031	149 + 461 + 456 + 189	20bp Pupps Promoter Fusion
424	**pANM-1_127**	pAGM8031	149 + 461 + 459 + 189	100bp Pupps Promoter Fusion
482	**pJBM-MC_281**	pSVM-mc	382 + 514 + 515 + 440	rasP complement

10 - Table 8.4 - Level M part library

MoClo-encoded level 1 parts either generated or used in this study. The parts of listed in numerical order. Internal name abbreviations: (pAS) generated by Andre Sim, (pJM) generated by Julia Manning, (pJB) generated by Jessica Bzdok, (pAN) generated by Annis Newman. The original name of MocClo destination vectors, in which the parts are encoded as well as the original names of the donor plasmids and primers used for PCR-amplification, or oligonucleotide annealing are indicated.

Additional plasmids				
ID	Plasmid	Resistance	Reference	Description
pJMP1	-	(E)Spc^R/(B)Cm^R	Peters et al Cell. 2016	Bacillus subtilis dCas9 expression vector; integrates into lacA/ganA
pJMP2	-	(E)Spc^R/(B)Cm^R	Peters et al Cell. 2016	Bacillus subtilis sgRNA expression vector; integrates into amyE

pDG1662	-	Amp^R	BGSC, Guerot-Fleury et al. 1996	ectopic integration into the *B. subtilis amyE* locus
GFE125	pDG1662	Amp^R	This study	pxylA+uppS>rasP for ectopic intergration into AmyE in B.Subtilis with cat resistance
GFE033	pJMP2	(E)Spc^R/(B)Cm^R	This study	rpsB sgRNA in pJMP2
GFE034	pJMP2	(E)Spc^R/(B)Cm^R	This study	tsf sgRNA in pJMP2
GFE035	pJMP2	(E)Spc^R/(B)Cm^R	This study	pyrH sgRNA in pJMP2
GFE036	pJMP2	(E)Spc^R/(B)Cm^R	This study	frr sgRNA in pJMP2
GFE037	pJMP2	(E)Spc^R/(B)Cm^R	This study	uppS sgRNA in pJMP2
GFE038	pJMP2	(E)Spc^R/(B)Cm^R	This study	cdsA sgRNA in pJMP2
GFE039	pJMP2	(E)Spc^R/(B)Cm^R	This study	dxr sgRNA in pJMP2
GFE040	pJMP2	(E)Spc^R/(B)Cm^R	This study	rasP sgRNA in pJMP2
GFE041	pJMP2	(E)Spc^R/(B)Cm^R	This study	proS sgRNA in pJMP2
GFE042	pJMP2	(E)Spc^R/(B)Cm^R	This study	polC sgRNA in pJMP2

11 - Table 8.5 – Additional used plasmids

Additional plasmids that were created or used in this study. The plasmids are listed in alphanumerical order. The internal id, plasmid backbone, resistance required for selection, source and description are provided. Cm^R: chloramphenicol resistance, Amp^R: ampicillin resistance, Spc^R: spectinomycin resistance. (E) – For selection in *Escherichia coli*, (B) – For section in *Bacillus subtilis*

Oligonucleotides		
Name	Nucleotide sequence (5' -> 3')	Description
GF0323	AGCTCGCAAATTAAACATCCC	rpsB_p2
GF0324	TGTTAATGTACCGCCCAACC	rpsb_p3
GF0326	CGTAGCGAAAATGGTTGAAGG	tsf_p2
GF0327	AGGGCTTGCTGAGTTAAGATTTG	tsf_p3
GF0329	CGTTGACGGTGTGTATAATGC	pyrH_p2
GF0330	AACGTCTGCCTCAATTTCAG	pyrH_p3
GF0332	ATCACGCCATACGATAAAACAG	frr_p2
GF0333	GATTTAAAGGTGTCTGCGCTCC	frr_p3
GF0335	TCTTTACTGACGTCTTGTGGC	upps_p2
GF0336	AATTACTCAGCCTTATCTCGCC	upps_p3
GF0338	CGCTGTTTTTGTCTGTTTTTGG	cdsA_p2
GF0339	AGATAGATATACGGGATCGGAAG	cdsA_p3
GF0341	TATCGAAAAGGCACTAACCCG	dxr_p2
GF0342	TTCGTTTGCCGCATTTAGC	dxr_p3
GF0344	TTGACGGAGGAAGACTGTTG	rasP_p2
GF0345	AACGCCGCAAACTGAAATAAG	rasP_p3
GF0347	CTTCCAATCCGCATCACTGTC	proS_p2
GF0348	AGAATATGAAGATCGTACGGCG	proS_p3
GF0382	tttagaagacatgcttAGGAATCTCATGCTCAACATACTC	upps HR fwd 1

GF0383	tttagaagacatGTCCTCTTTTGTATAACGTTCTAAGT	upps HR rev 1
GF0384	tttagaagacatGGACATACTTAAGGGAGAAATTCCC	upps HR fwd 2
GF0385	tttagaagacatagcgTGACAATTTCAGTACGGCCTC	upps HR rev 2
GF0386	tttagaagacatggagGGGGAAATAACGTGTCAAAAGAAG	frr HR fwd 1
GF0387	tttagaagacatGATGACAACTGATTTAAAGGTGTCT	frr HR rev 1
GF0388	tttagaagacatCATCTATTAACGTGCCTGAAGC	frr HR fwd 2
GF0389	tttagaagacatGTCCTCAGTGGAAGCACGC	frr HR rev 2
GF0390	tttagaagacatGGACGTTCAAAAACTGACAGATG	frr HR fwd 3
GF0391	tttagaagacatcattCCCTGCTGATAATCAATGTAATCA	frr HR rev 3
GF0392	ggagTTAATACTGTTGATTACATTGATTATCAGCAGGGAATGTAACCTTTTTGGGTGACGG	upps promoter
GF0393	agtaCCGTCACCCAAAAAGGTTACATTCCCTGCTGATAATCAATGTAATCAACAGTATTAA	upps promoter 2
GF0481	GACCACGCGTATCGATGTCGACTTTTTTTTTTTTTTTTC	dT-RACE-Anchor
GF0482	GACCACGCGTATCGATGTCGAC	RACE-Adaptor
GF0483	ACGTACCGTTTTCTTGCATT	rpSB_p5
GF0484	TCGCAGTTTCTTCATCAGTTG	tsf_p4
GF0544	tttagaagacat*ggag*ACTATATGGGAATGCTGGCGAC	frr HR fwd new
GF0545	tttagaagacat**A**TCTTCAATATTCAAAGAGCCTTCC	upps HR rev 2 new
GF0546	tttagaagacatAGA**T**ATTGATGAATCGCTTTTTTCTAC	upps HR fwd 3 new
GF0547	tttagaagacat*agcg*TTTTGGAAATCCCGTCCGCTTC	upps HR rev 3 new
GF0561	ACATTTATTGTACAACACGAGCCCATTTTTG	Universal reverse sgRNA primer
GF0562	gcttcatttttgggttccaaGTTTTAGAGCTAGAAATAGCAAGTTAAAATAAGGC	rpsB sgRNA
GF0563	tagaaccttctgctgcgataGTTTTAGAGCTAGAAATAGCAAGTTAAAATAAGGC	tsf sgRNA
GF0564	aatacgatacgtttgtatttGTTTTAGAGCTAGAAATAGCAAGTTAAAATAAGGC	pyrH sgRNA
GF0565	ctaataatgatggatttgcaGTTTTAGAGCTAGAAATAGCAAGTTAAAATAAGGC	frr sgRNA
GF0566	ttttgtataacgttctaagtGTTTTAGAGCTAGAAATAGCAAGTTAAAATAAGGC	uppS sgRNA
GF0567	gcataaattaatatggtgaaGTTTTAGAGCTAGAAATAGCAAGTTAAAATAAGGC	cdsA sgRNA
GF0568	acatagataccagctgaaatGTTTTAGAGCTAGAAATAGCAAGTTAAAATAAGGC	dxr sgRNA
GF0569	cggcatcagctggaacttcaGTTTTAGAGCTAGAAATAGCAAGTTAAAATAAGGC	proS sgRNA
GF0570	atgttcaaagtatgtcatgaGTTTTAGAGCTAGAAATAGCAAGTTAAAATAAGGC	polC sgRNA
GF0607	tttagaagacatggagGGTGTGACTGAAGAACAGATTGG	post-polC HR left fwd
GF0608	tttagaagacatcattTTACGATGGCACTTTTTGCG	post-polC HR left rev
GF0609	tttagaagacatgcttGAGGCAAAGAGTGGGGAAACC	post-polC HR right fwd
GF0610	tttagaagacatagcgTATGCGCCCGAAATCCTTAG	post-polC HR right rev
GF0617	tttagaagacataatgATATCGACCCAAGTACCGCC	tet array fwd 2

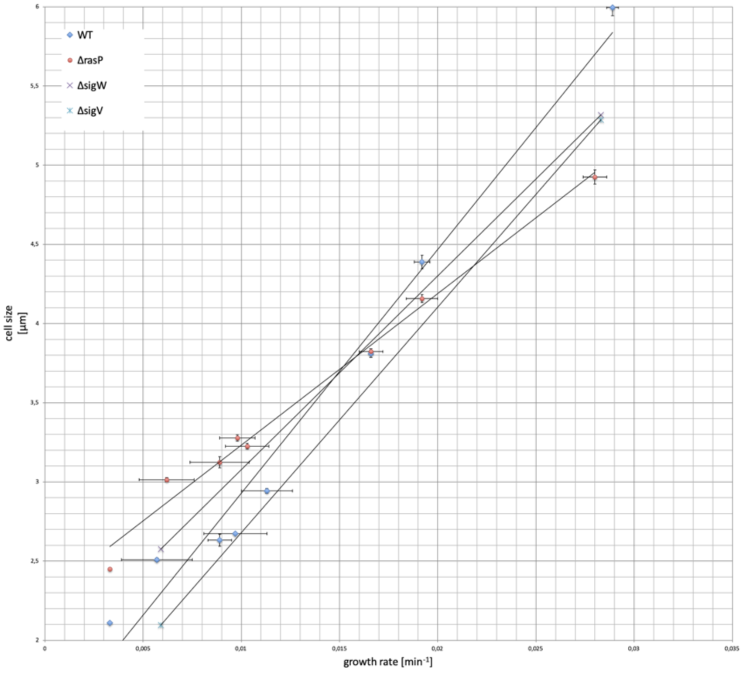
GF0618	tttagaagacataagcTGATAGGGACAGCGCTGAGT	tet array rev 2
GF0627	ATCACCATGGACAAGCACAA	qPCR mnaA fwd
GF0628	TCACATCAAGCCTGACTTCG	qPCR mnaA rev
GF0629	GACGGCATTACGGTTGAAGT	qPCR gyrB fwd
GF0630	CCGCCTTCGTACGTGTTAAT	qPCR gyrB rev
GF0631	GTTCGGCAAAGGTTCCATTA	qPCR recA fwd
GF0632	GCCAATTCCCAGTGCTGTAT	qPCR recA rev
GF0645	CGGAAGCTCGCAAATTAAAC	qPCR rpsB fwd
GF0646	CAACTTCCGCTTCTTCTTCG	qPCR rpsB rev
GF0647	AAACTGCGGCTTCCAACTTA	qPCR uppS fwd
GF0648	CACCTTCATCCCTTCATGGT	qPCR uppS rev
GF0649	CTTGGGATTGTCAACCTGCT	qPCR rasP fwd
GF0650	TTCCATGTGACAACCAGCAT	qPCR rasP rev
GF0688	TGTACTGGGGTGGATGCAG	tetR-YFP fwd for pGP380
GF0689	TTTGGATCCAAACCACTTCGTGCAGAAGAC	tetR-YFP rev for pGP380
GF0702	tttagaagacatgcttGCATTTGTTGTGTTTATCGGAGTAG	after rasP HR fwd
GF0703	tttagaagacatAGCGCCGAAACCTGTAAGAACGC	after rasP HR rev
GF0705	tttagaagacat**CATT**AGATTCCTCCGTCACCCAAA	Universal uppS promoter rev
GF0706	**GGAG**TTGGGTGACGGAGGAATCT	20bp uppS promoter fwd
GF0707	**CATT**AGATTCCTCCGTCACCCAA	20bp uppS promoter rev
GF0708	**GGAG**CAGCAGGGAATGTAACCTTTTTGGGTGACGGAGGAATCT	40bp uppS promoter fwd
GF0709	**CATT**AGATTCCTCCGTCACCCAAAAAGGTTACATTCCCTGCTG	40bp uppS promoter rev
GF0710	**GGAG**CTGTTGATTACATTGATTATCAGCAGGGAATGTAACCTTTTTGGGTGACGGAGGAATCT	60bp uppS promoter fwd
GF0711	**CATT**AGATTCCTCCGTCACCCAAAAAGGTTACATTCCCTGCTGATAATCAATGTAATCAACAG	60bp uppS promoter rev
GF0712	tttagaagacat**GGAG**AGGGGGTTTTTTGTTAATACTGTTG	80bp uppS promoter fwd
GF0713	tttagaagacat**GGAG**AAAGACCCTCTCATGTTTACAGG	100bp uppS promoter fwd
GF0714	tttagaagacat**GGAG**TGTACAATAGATAATAGTGAAAAGACCCTC	120bp uppS promoter fwd
GF0715	tttagaagacat**GGAG**GGAAGTTTAATGAAAAACTATGTACAATAGATAATAGTGA	140bp uppS promoter fwd
GF0716	tttagaagacat**GGAG**GACAAAGAAAAAGAAATCATGGAAGT	160bp uppS promoter fwd
GF0717	tttagaagacat**GGAG**AAATTGACAGTGTCACAAAAGACAA	180bp uppS promoter fwd
GF0718	tttagaagacat**GGAG**GACAGATGAATATGTGTCAAAAATTGAC	200bp uppS promoter fwd
GF0719	tttagaagacat**GGAG**ACTGAAGACGTTCAAAAACTGAC	220bp uppS promoter fwd

GF0720	tttagaagacat**GGAG**AGGATGAACTGCGTGCTT	240bp uppS promoter fwd
GF0721	tttagaagacat**GGAG**GAAAAACGGAGACATTACTGAGG	260bp uppS promoter fwd
GF0722	tttagaagacat**GGAG**GATGATCTCAAAAAACTTGAGAAAAACGG	280bp uppS promoter fwd
GF0723	tttagaagacat**GGAG**ACGTTCGCCGTGATGCTAA	300bp uppS promoter fwd
GF0759	tttagaagacatGGAGAATATGATGTTCCGCTGCTG	rasP HR Left fwd 1
GF0760	tttagaagacatG**C**CTTCAATAGCCAAAAACGGT	rasP HR Left rev 1
GF0761	tttagaagacatAG**G**CTGTATCGAAAAGGCAC	rasP HR Left fwd 2
GF0762	tttagaagacatCATTCATGGAAGAAAACGAGCG	rasP HR Left rev 2
GF0765	**GGAG**AAATTGACAGTGTCACAAAAGACAAAGAAAAAGAAATCATGGAAGTTTAAT	PuppS no UTR FWD
GF0766	**AGTA**ATTAAACTTCCATGATTTCTTTTTCTTTGTCTTTTGTGACACTGTCAATTT	PuppS no UTR REV
GF0793	TGCTCGTTATCACGCCATAC	qPCR frr
GF0794	TCGAATCATATTGCCGTCAC	qPCR frr
GF0795	GGAGCATCATTTCACGGAGT	qPCR lux
GF0796	GGGCTGTGGGAAGAACAATA	qPCR lux
GF0802	GACAGAATACGAGCGGGGTA	qPCR proS F
GF0803	CGGCATCAATTTTCTCCATT	qPCR proS R
GF0842	tttagaagacatTCTCCCTCCGTCAAGTGCC	rasP-gene rev1
GF0843	tttagaagacatGAGACTGTTGTTTCTATTTATTGAAGCG	rasP-gene fwd2
GF0844	tttagaagacataagcCTTTTCGTTTACAAAAACAGCCGC	rasP-gene rev2
GF0850	tttagaagacataatgTTCGTGAATACAGTTATAGCGTTTATCAT	rasP-gene fwd1

12 - Table 8.6 – Oligonucleotides used in this study

The list of oligonucleotides used in this study sorted in numerical order. Nucleotide sequences are in 5’ -> 3’ order.

8.2 Supplementary



69 - Figure 8.1 - rasP knockout mutants are smaller than wild type in rich media but smaller in poor media

The effect of *rasP, sigW and sigV* deletions on cell length in *Bacillus subtilis* at different growth rates. Each data point represents cells from different strains and media. Vertical error bars represent the standard error in the cell length, horizontal error bars represent the standard error in the growth rate. Linear regressions were fit to the data points each strain type. In order from highest nutritional quality to lowest, LB media, MOPS media + glucose and amino acids (aa), MOPS media + glucose, MOPS media, + glycerol and aa, MOPS media + glycerol, MOPS media + ribose and aa, and MOPS media + ribose.

Gene clusters identified at the Bacteria taxonomic level

Cluster ID	Gene Descriptions
1	K Homology domain, 30S ribosomal protein S10, 50S ribosomal protein L23, Ribosomal protein L2, 30S ribosomal protein S19, 50S ribosomal protein L16, 30S ribosomal protein S17, 50S ribosomal protein L14, 50S ribosomal protein L24, 50S ribosomal protein L18, 50S ribosomal protein L30, 50S ribosomal protein L15, Protein translocase subunit SecY, translation initiation factor IF-1, 30S ribosomal protein S11, 30S ribosomal protein S4, 50S ribosomal protein L4, 50S ribosomal protein L3, 50S ribosomal protein L22, 50S ribosomal protein L29, 50S ribosomal protein L5, 30S ribosomal protein S14, 30S ribosomal protein S8, 50S ribosomal protein L6, 30S ribosomal protein S5, adenylate kinase, 30S ribosomal protein S13, Methionine aminopeptidase, DNA-directed RNA polymerase subunit alpha, 50S ribosomal protein L17, 50S ribosomal protein L36, 30S ribosomal protein S12, preprotein translocase subunit SecE, Transcription termination/antitermination protein NusG, 50S ribosomal protein L11, 50S ribosomal protein L1, 50S ribosomal protein L10, 50S ribosomal protein L7/L12, DNA-directed RNA polymerase subunit beta, DNA-directed RNA polymerase subunit beta', 30S ribosomal protein S7, elongation factor G, Elongation factor Tu
2	glucose-1-phosphate thymidylyltransferase, dTDP-glucose 4,6-dehydratase, dTDP-4-dehydrorhamnose reductase, dTDP-4-dehydrorhamnose 3,5-epimerase, glycosyl transferase, NAD-dependent epimerase, glycosyl transferase, tyrosine protein kinase, sugar ABC transporter substrate-binding protein, undecaprenyl-phosphate glucose phosphotransferase, acetyltransferase, UDP-4-amino-4, 6-dideoxy-N-acetyl-beta-L-altrosamine transaminase, UDP-glucose 4-epimerase, glycosyltransferase WbuB, glycosyl transferase family 1, glycosyl transferase family 2, polysaccharide biosynthesis protein, glycosyl transferase family 2, glucose-1-phosphate adenylyltransferase, glycosyl transferase, Transport permease protein, ABC transporter ATP-binding protein
3	NADH-quinone oxidoreductase subunit E, NADH-ubiquinone oxidoreductase 51kDa subunit, FMN-binding domain, NADH-quinone oxidoreductase, NADH dehydrogenase, NADH dehydrogenase, NADH-quinone oxidoreductase subunit D, NADH-quinone oxidoreductase subunit H, NADH:ubiquinone oxidoreductase subunit J, NADH-quinone oxidoreductase subunit K, NADH-quinone oxidoreductase subunit M, NADH:ubiquinone oxidoreductase subunit N, 4Fe-4S ferredoxin, iron-sulphur binding, conserved site, NADH-quinone oxidoreductase subunit A, NADH, Ion-translocating oxidoreductase complex subunit A, Ion-translocating oxidoreductase complex subunit D, Na(+)-translocating NADH-quinone reductase subunit D
4	cell division protein FtsZ, 16S rRNA (cytosine(1402)-N(4))-methyltransferase, penicillin-binding protein, UDP-N-acetylmuramoyl-L-alanyl-D-glutamate--2, 6-diaminopimelate ligase, Phospho-N-acetylmuramoyl-pentapeptide-transferase, cell division protein FtsW, rod shape-determining protein, penicillin-binding protein 2, rod shape-determining protein MreC, UDP-N-acetylmuramate--L-alanine ligase, Cell division protein FtsQ, cell division protein FtsA, UDP-N-acetylglucosamine--N-acetylmuramyl- (pentapeptide) pyrophosphoryl-undecaprenol N-acetylglucosamine transferase, Transcriptional regulator MraZ, UDP-N-acetylmuramoylalanine-D-glutamate ligase, UDP-3-O-acyl-N-acetylglucosamine deacetylase
5	Beta sliding clamp, chromosome partitioning protein ParB, tRNA uridine 5-carboxymethylaminomethyl modification enzyme MnmG, Ribosomal RNA small subunit methyltransferase G, chromosome partitioning protein, GTPase Der, Membrane protein insertase YidC, DNA gyrase subunit B, DNA gyrase subunit A, Chromosomal replication initiator protein DnaA, DNA replication and repair protein RecF, 50S ribosomal protein L34, ribonuclease P protein component
6	chromosome segregation protein SMC, signal recognition particle protein, 30S ribosomal protein S16, Ribosome maturation factor RimM, tRNA (guanine(37)-N(1))-methyltransferase, 50S ribosomal protein L19, RNA-binding protein, ribonuclease III, Signal peptidase I
7	ATP synthase epsilon chain, ATP synthase subunit c, ATP synthase subunit delta, ATP synthase subunit alpha, ATP synthase gamma chain, ATP synthase subunit beta, ATP synthase subunit a, ATP synthase subunit b

8	Histidinol-phosphate aminotransferase, 1-(5-phosphoribosyl)-5-, histidinol dehydrogenase, ATP phosphoribosyltransferase, imidazoleglycerol-phosphate dehydratase, imidazole glycerol phosphate synthase subunit HisH, imidazole glycerol phosphate synthase cyclase subunit, phosphoribosyl-AMP cyclohydrolase
9	ABC transporter ATP-binding protein, ABC transporter permease, Efflux transporter, RND family, MFP subunit, acriflavin resistance protein, RND transporter, efflux transporter periplasmic adaptor subunit, Histidine kinase, DNA-binding response regulator
10	Uridylate kinase, 30S ribosomal protein S2, Elongation factor Ts, ribosome recycling factor, di-trans,poly-cis-decaprenylcistransferase, phosphatidate cytidylyltransferase, 1-deoxy-D-xylulose-5-phosphate reductoisomerase, Zinc metalloprotease
11	3-isopropylmalate dehydratase small subunit, 3-isopropylmalate dehydratase large subunit, 3-isopropylmalate dehydrogenase, 2-isopropylmalate synthase, acetolactate synthase small subunit, ketol-acid reductoisomerase, Acetolactate synthase
12	chemotaxis protein CheR, chemotaxis response regulator protein-glutamate methylesterase, chemotaxis protein CheA, response regulator, chemotaxis protein CheW, methyl-accepting chemotaxis protein, response regulator
13	anthranilate phosphoribosyltransferase, anthranilate synthase component I, Indole-3-glycerol phosphate synthase, Tryptophan synthase beta chain, Tryptophan synthase alpha chain, N-(5'-phosphoribosyl)anthranilate isomerase, glutamine amidotransferase
14	tRNA pseudouridine synthase B, Ribosome maturation factor RimP, Transcription termination/antitermination protein NusA, ribosome-binding factor A, 30S ribosomal protein S15, Polyribonucleotide nucleotidyltransferase, translation initiation factor IF-2
15	flagellar biosynthetic protein FliP, Flagellar biosynthetic protein FliQ, Flagellar biosynthetic protein FliR, Flagellar motor switch protein FliN, Flagellar biosynthetic protein FlhB, flagellar biosynthesis protein FlhA
16	50S ribosomal protein L20, threonine--tRNA ligase, 50S ribosomal protein L35, phenylalanine--tRNA ligase subunit alpha, phenylalanine--tRNA ligase subunit beta
17	monovalent cation/H+ antiporter subunit D, Na+/H+ antiporter subunit C, Na+/H+ antiporter subunit E, K+/H+ antiporter subunit F, Na+/H+ antiporter subunit G
18	glycine cleavage system protein H, glycine cleavage system protein T, glycine dehydrogenase (aminomethyl-transferring), FAD-dependent oxidoreductase, pyridoxamine 5'-phosphate oxidase
19	replicative DNA helicase, 30S ribosomal protein S6, single-stranded DNA-binding protein, 30S ribosomal protein S18, 50S ribosomal protein L9
20	ABC transporter permease, ABC transporter, nitrate ABC transporter substrate-binding protein, ABC transporter permease, ABC transporter permease
21	pyruvate dehydrogenase (acetyl-transferring) E1 component subunit alpha, transketolase, dihydrolipoamide succinyltransferase, dihydrolipoyl dehydrogenase, 2-oxoglutarate dehydrogenase E1 component
22	phosphate ABC transporter substrate-binding protein PstS, phosphate ABC transporter permease subunit PstC, phosphate ABC transporter, permease protein PstA, phosphate ABC transporter ATP-binding protein, phosphate transport system regulatory protein PhoU
23	amino acid ABC transporter ATP-binding protein, amino acid ABC transporter permease, amino acid ABC transporter substrate-binding protein, amino acid ABC transporter substrate-binding protein, amino acid ABC transporter substrate-binding protein
24	Uroporphyrinogen-III synthase, sulfite reductase, phosphoadenosine phosphosulfate reductase, precorrin-6Y C5,15-methyltransferase, precorrin-2 C(20)-methyltransferase
25	hemolysin D, cation transporter, ABC transporter, ABC transporter, Transport permease protein
26	ATP-dependent protease subunit HslV, ATP-dependent Clp protease ATP-binding subunit ClpX, trigger factor, ATP-dependent Clp protease proteolytic subunit, Lon protease
27	peptide ABC transporter ATP-binding protein, peptide ABC transporter ATP-binding protein, peptide ABC transporter substrate-binding protein, ABC transporter permease, peptide ABC transporter permease
28	iron ABC transporter ATP-binding protein, iron ABC transporter permease, ABC transporter, ABC transporter substrate-binding protein, ABC transporter substrate-binding protein
29	ATPase AAA, membrane protein, VWA domain-containing protein, cobaltochelatase subunit CobS
30	flagellar basal body rod protein FlgC, Flagellar basal body rod protein FlgB, Flagellar hook protein FlgE, flagellar hook capping protein
31	iron-sulfur cluster assembly scaffold protein, cysteine desulfurase, Rrf2 family transcriptional regulator, iron-sulfur cluster insertion protein ErpA
32	flagellar protein FliS, Flagellin, flagellar hook-associated protein 3, Flagellar hook-associated protein 1
33	Heat-inducible transcription repressor HrcA, Chaperone protein DnaJ, Protein GrpE, Chaperone protein DnaK
34	cytochrome o ubiquinol oxidase subunit III, cytochrome c oxidase subunit II, cytochrome c oxidase subunit I, protoheme IX farnesyltransferase
35	glutamate 5-kinase, Ribosome-binding ATPase YchF, 50S ribosomal protein L21, 50S ribosomal protein L27
36	transcriptional regulator, crossover junction endodeoxyribonuclease RuvC, Holliday junction DNA helicase RuvA, Holliday junction DNA helicase RuvB
37	riboflavin synthase subunit alpha, 3,4-dihydroxy-2-butanone 4-phosphate synthase, 6,7-dimethyl-8-ribityllumazine synthase, Riboflavin biosynthesis protein RibD
38	AraC family transcriptional regulator, TonB-dependent receptor, RNA polymerase sigma factor, anti-FecI sigma factor FecR
39	30S ribosomal protein S1, cytidylate kinase, 3-phosphoshikimate 1-carboxyvinyltransferase, DNA-binding protein
40	ABC transporter ATP-binding protein, SUF system FeS cluster assembly, SufBD, SUF system FeS cluster assembly, SufBD, Cysteine desulfurase
41	xanthine dehydrogenase small subunit, xanthine dehydrogenase molybdopterin binding subunit, xanthine dehydrogenase accessory protein XdhC, (2Fe-2S)-binding protein
42	Type I restriction enzyme R Protein, N6 adenine-specific DNA methyltransferase, N-terminal domain, restriction endonuclease subunit S
43	C4-dicarboxylate ABC transporter permease, TRAP transporter solute receptor DctP superfamily, C4-dicarboxylate ABC transporter permease
44	recombination protein RecR, DNA polymerase III subunit gamma/tau, nucleoid-associated protein
45	Thymidylate kinase, DNA polymerase III subunit delta', hydrolase TatD
46	Flagellar hook-basal body complex protein FliE, flagellar motor switch protein FliG, Flagellar M-ring protein
47	membrane protein, metal ABC transporter substrate-binding protein, ABC transporter
48	1-phosphofructokinase, PTS fructose transporter subunit IIC, DeoR family transcriptional regulator
49	Triosephosphate isomerase, Glyceraldehyde-3-phosphate dehydrogenase, phosphoglycerate kinase

50	ABC transporter substrate-binding protein, C4-dicarboxylate ABC transporter permease, C4-dicarboxylate ABC transporter
51	amidase, Aspartyl/glutamyl-tRNA(Asn/Gln) amidotransferase subunit C, Aspartyl/glutamyl-tRNA(Asn/Gln) amidotransferase subunit B
52	methylcrotonoyl-CoA carboxylase, acetyl-CoA carboxylase biotin carboxylase subunit, Biotin carboxyl carrier protein of acetyl-CoA carboxylase
53	ABC transporter ATP-binding protein, ABC transporter permease, glycine/betaine ABC transporter substrate-binding protein
54	Protein translocase subunit SecD, preprotein translocase subunit YajC, Protein-export membrane protein SecF
55	GTP 3',8-cyclase, molybdenum cofactor biosynthesis protein, molybdopterin molybdenumtransferase MoeA
56	membrane protein, phosphate starvation protein PhoH, Endoribonuclease YbeY
57	Ribose-phosphate pyrophosphokinase, 50S ribosomal protein L25, Peptidyl-tRNA hydrolase

13 - Table 8.7 – Identified gene clusters from the Bacteria taxonomic level

A list of the clusters that were identified at the Bacteria taxonomic level (the highest). Clusters are sorted in descending order by the number of genes. The descriptions column contain the OrthoDB assigned gene description for each gene member in the cluster.

Horizontal gene transfer gene clusters	
Cluster ID	Gene Descriptions
1	multidrug transporter, glycosyl transferase, hemolysin D, alginate biosynthesis protein AlgE, poly(beta-D-mannuronate) lyase
2	dialkylrecorsinol condensing enzyme DarA, F3YYE9_DESAF, peptidase, 3-oxoacyl-ACP synthase, 3-oxoacyl-ACP synthase
3	ABC transporter ATP-binding protein, ABC transporter ATP-binding protein, branched-chain amino acid ABC transporter substrate-binding protein, branched-chain amino acid ABC transporter permease, branched-chain amino acid ABC transporter permease
4	L1L958_9ACTN, Ferritin-like, Secreted protein, membrane protein
5	peptidase, Signal peptide protein, histidine kinase, kinesin
6	type II secretion protein F, Conserved domain protein, P-loop containing nucleoside triphosphate hydrolase, type II secretion system protein F
7	microcompartment protein EutL, ethanolamine utilization protein EutH, ethanolamine ammonia-lyase, ethanolamine utilization cobalamin adenosyltransferase
8	A0A101PM64_9ACTN, lytic transglycosylase, periplasmic immunogenic protein, Homing endonuclease
9	citrate lyase subunit alpha, Citrate lyase acyl carrier protein, holo-ACP synthase CitX, \N
10	glucose-1-phosphate thymidylyltransferase, dTDP-glucose 4,6-dehydratase, dTDP-4-dehydrorhamnose reductase, dTDP-4-dehydrorhamnose 3,5-epimerase
11	phosphoesterase, phosphoesterase, type IV secretion protein IcmB
12	A0A126V038_9RHOB, threonine--tRNA ligase, V9WKJ0_9RHOB
13	DUF1127 domain-containing protein, Holliday junction resolvasome, helicase subunit, Domain of unknown function DUF4105
14	ribonuclease P protein component, 50S ribosomal protein L34, Membrane protein insertase YidC
15	phage tail protein, H2K0L2_STRHJ, A0A0K2AYC9_STRAM
16	P-loop containing nucleoside triphosphate hydrolase, acetolactate synthase, Imidazole glycerol phosphate synthase subunit HisH
17	membrane protein, DNA replication terminus site-binding protein, nucleoid-associated protein
18	Protein of unknown function DUF1896, Protein of unknown function DUF3945, tetracycline regulation of excision, RteC
19	prepilin-type N-terminal cleavage/methylation domain-containing protein, prepilin-type N-terminal cleavage/methylation domain-containing protein, prepilin-type N-terminal cleavage/methylation domain-containing protein
20	phosphoribosyltransferase, membrane protein, Short C-terminal domain protein
21	prokaryotic E2 family E, A0A0U3JEM4_ACIJO, UBA/THIF-type NAD/FAD binding protein
22	HNH endonuclease, Secreted protein, A0A1Q5MCT3_9ACTN
23	flagellar biosynthesis repressor FlbT, flagellar biosynthesis regulator FlhF, flagellar protein FlgJ
24	F2K9A2_PSEBN, lipoprotein, terminase
25	Gll3097 protein, Photosystem II CP47 reaction center protein, Photosystem II reaction center protein T
26	membrane protein, Tsr2248 protein, CP12 domain protein
27	A0A0N9WZW2_PSEFL, lipoprotein, F2KLH6_PSEBN
28	membrane protein, membrane protein, membrane protein
29	Armadillo-type fold, 3-oxoacyl-ACP synthase, Uncharacterized protein conserved in bacteria
30	transporter, membrane protein, membrane protein
31	Putative membrane protein, membrane protein, Hypothetical membrane protein
32	DUF3153 domain-containing protein, polyketide cyclase / dehydrase and lipid transport, I7GBP5_MYCS2
33	D-proline reductase (dithiol) proprotein PrdA, C8P4I6_9LACO, permease
34	stage 0 sporulation protein, Nitrogen regulatory PII-like, alpha/beta, Initiation-control protein YabA
35	sensory transduction regulator, membrane protein, lipoprotein
36	type VI secretion protein, Type IV secretion system, VirB5, type IV secretion system protein VirB10
37	A0A1D8G7F0_9ACTN, membrane protein, membrane protein
38	pilus assembly protein TadE, pilus assembly protein TadG, pilus assembly protein TadG
39	TerD-family protein, A0A1M7Q688_9ACTN, DUF4937 domain-containing protein
40	A4WRX2_RHOS5, primosomal protein DnaI, A0A0B5DS84_9RHOB
41	A0A0X3RV14_STRRM, A0A191V936_9ACTN, A0A0N0SUC9_9ACTN
42	polyketide cyclase, polyketide cyclase, cupin
43	Chain length determinant family protein, uracil phosphoribosyltransferase, membrane protein
44	molecular chaperone DnaJ, nosiheptide resistance regulatory protein, DNA-binding protein
45	sugar ABC transporter substrate-binding protein, sugar ABC transporter permease, ABC transporter permease

46	lipid-A-disaccharide synthase, Acyl-, outer membrane protein assembly factor BamA
47	ABC transporter permease, sugar ABC transporter substrate-binding protein, sugar ABC transporter ATP-binding protein
48	ABC transporter permease, BMP family ABC transporter substrate-binding protein, ABC transporter permease
49	Tetratricopeptide repeat, pilus assembly protein, ATPase
50	ATP synthase subunit D, Protein of unknown function DUF2764, ATP synthase subunit E
51	stage VI sporulation protein D, membrane protein, membrane protein
52	lipoprotein, A0A1P8MW48_9RHOB, A4WVE9_RHOS5
53	Secreted protein, Secreted protein, cobalt ABC transporter permease
54	Protein of unknown function DUF1419, L0LRA9_RHITR, Uncharacterised conserved protein UCP036055
55	permease, Chromosome segregation ATPase-like protein, non-ribosomal peptide synthetase module
56	sugar ABC transporter permease, ABC transporter substrate-binding protein, ABC transporter permease
57	ABC transporter ATP-binding protein, mammalian cell entry protein, ABC transporter permease
58	Beta sliding clamp, Chromosomal replication initiator protein DnaA, DNA replication and repair protein RecF

14 - Table 8.8 - Horizontal gene transfer gene clusters

A list of the clusters that were identified by normalizing against ortholog distribution instead of taxa size. These clusters were compared the original taxa size normalized cluster list, leaving on the unique clusters in this list. Clusters are sorted in descending order by the number of genes. The descriptions column contain the OrthoDB assigned gene description for each gene member in the cluster.

RNA datasets			
Species	Dataset	Experiment	Comments
Bacillus subtilis	SRP022234	SRR899517	Control
		SRR899518	Control
		SRR899519	Acetic Acid
		SRR899520	Acetic Acid
		SRR899521	Ethanol
		SRR899535	Ethanol
		SRR899548	Lactic Acid
		SRR899549	Lactic Acid
		SRR899550	Indole
		SRR899551	Indole
		SRR899552	Low H2O2
		SRR899553	Low H2O2
		SRR899554	High H2O2
		SRR899555	High H2O2
		SRR899556	oxDXS
		SRR899557	oxDXS
		SRR899558	oxFNI
		SRR899559	oxFNI
		SRR899645	DMSO
		SRR899646	DMSO
		SRR899649	oxlspA
		SRR899650	oxlspA
		SRR922367	control
		SRR922368	low acetic acid
		SRR922369	High Acetic Acid
		SRR922370	Low Ethanol
		SRR922371	High Ethanol
		SRR922372	Low Lactic Acid
		SRR922373	High Lactic Acid
		SRR922374	Low Indole
		SRR922375	High Indole
		SRR922376	Low H2O2
		SRR922377	High H2O2
		SRR922378	NaCl
		SRR922379	Glucose
		SRR922380	Mannose
		SRR922381	Xylose
		SRR922382	oxDXS
		SRR922383	oxDXSDXR
		SRR922384	oxDXSFNI
		SRR922385	oxFNI
	SRP074602	SRR3488622.sra	GSM2147025: salt.T90 (KN14_R1); Bacillus subtilis subsp. subtilis str. 168; RNA-Seq
		SRR3488623.sra	GSM2147024: salt.T60 (KN13_R1); Bacillus subtilis subsp. subtilis str. 168; RNA-Seq
		SRR3488624.sra	GSM2147023: no_salt.T90 (KN3_R2); Bacillus subtilis subsp. subtilis str. 168; RNA-Seq
		SRR3488625.sra	GSM2147022: no_salt.T60 (KN2_R2); Bacillus subtilis subsp. subtilis str. 168; RNA-Seq

		SRR3488626.sra	GSM2147021: no_salt.T30 (KN1_R2); Bacillus subtilis subsp. subtilis str. 168; RNA-Seq
		SRR3488627.sra	GSM2147020: no_salt.T90 (KN11); Bacillus subtilis subsp. subtilis str. 168; RNA-Seq
		SRR3488628.sra	GSM2147019: no_salt.T60 (KN10); Bacillus subtilis subsp. subtilis str. 168; RNA-Seq
		SRR3488629.sra	GSM2147018: no_salt.T30 (KN9); Bacillus subtilis subsp. subtilis str. 168; RNA-Seq
		SRR3488630.sra	GSM2147017: dormant.T0 (KN8); Bacillus subtilis subsp. subtilis str. 168; RNA-Seq
		SRR3488631.sra	GSM2147016: salt.T90 (KN7); Bacillus subtilis subsp. subtilis str. 168; RNA-Seq
		SRR3488632.sra	GSM2147015: salt.T60 (KN6); Bacillus subtilis subsp. subtilis str. 168; RNA-Seq
		SRR3488633.sra	GSM2147014: salt.T30 (KN5); Bacillus subtilis subsp. subtilis str. 168; RNA-Seq
		SRR3488634.sra	GSM2147013: dormant.T0 (KN4); Bacillus subtilis subsp. subtilis str. 168; RNA-Seq
		SRR3488635.sra	GSM2147012: salt.T30 (KN12_R1); Bacillus subtilis subsp. subtilis str. 168; RNA-Seq
	SRP068910	SRR3124507.sra	PARE rnc+ ^rnjA
		SRR3124508.sra	PARE ^ rnc ^ rnjA
		SRR3124509.sra	P sac -rnjA ^rny PARE
		SRR3124510.sra	RNASeq +rnc rep I
		SRR3124511.sra	RNASeq +rnc rep II
		SRR3124512.sra	RNAsec ^rnc rep I
		SRR3124513.sra	RNAsec ^rnc rep II
Bacteroides fragilis	SRP063781	SRR2584315.sra	Bacteroides fragilis grown on glucose- replicate 2
		SRR2584332.sra	Bacteroides fragilis grown on glucose- replicate 3
		SRR2584350.sra	Bacteroides fragilis grown on glucose- replicate 1
		SRR2584375.sra	Bacteroides fragilis grown on mucin O-linked glycans- replicate 3
		SRR2585000.sra	Bacteroides fragilis grown on mucin O-linked glycans- replicate 2
		SRR2602448.sra	Bacteroides fragilis grown on mucin O-linked glycans-replicate 1
		SRR2827509.sra	GSM1919133: mRNA18_Mxn3, Mid log phase, Xylan carbon source; Bacteroides xylanisolvens; RNA-Seq
		SRR2827510.sra	GSM1919132: mRNA17_Mxn2, Mid log phase, Xylan carbon source; Bacteroides xylanisolvens; RNA-Seq
		SRR2827511.sra	GSM1919131: mRNA16_Mxn1, Mid log phase, Xylan carbon source; Bacteroides xylanisolvens; RNA-Seq
		SRR2827512.sra	GSM1919130: mRNA21_Mcp3, Mid log phase, Citrus Pectin carbon source; Bacteroides xylanisolvens; RNA-Seq
		SRR2827513.sra	GSM1919129: mRNA20_Mcp2, Mid log phase, Citrus Pectin carbon source; Bacteroides xylanisolvens; RNA-Seq
		SRR2827514.sra	GSM1919128: mRNA19_Mcp1, Mid log phase, Citrus Pectin carbon source; Bacteroides xylanisolvens; RNA-Seq
		SRR2827515.sra	GSM1919127: mRNA24_MG3, Mid log phase, Glucose carbon source; Bacteroides xylanisolvens; RNA-Seq
		SRR2827516.sra	GSM1919126: mRNA23_MG2, Mid log phase, Glucose carbon source; Bacteroides xylanisolvens; RNA-Seq
		SRR2827517.sra	GSM1919125: mRNA22_MG1, Mid log phase, Glucose carbon source; Bacteroides xylanisolvens; RNA-Seq
		SRR2827518.sra	GSM1919124: mRNA15, Late log phase, Xylose carbon source; Bacteroides xylanisolvens; RNA-Seq
		SRR2827519.sra	GSM1919123: mRNA14, Late log phase, Xylose carbon source; Bacteroides xylanisolvens; RNA-Seq
		SRR2827520.sra	GSM1919122: mRNA13, Late log phase, Xylose carbon source; Bacteroides xylanisolvens; RNA-Seq
		SRR2827521.sra	GSM1919121: mRNA3, Late log phase, Xylan carbon source; Bacteroides xylanisolvens; RNA-Seq
		SRR2827522.sra	GSM1919120: mRNA2, Late log phase, Xylan carbon source; Bacteroides xylanisolvens; RNA-Seq
		SRR2827523.sra	GSM1919119: mRNA1, Late log phase, Xylan carbon source; Bacteroides xylanisolvens; RNA-Seq
		SRR2827524.sra	GSM1919118: mRNA12, Late log phase, Glucose carbon source; Bacteroides xylanisolvens; RNA-Seq
		SRR2827525.sra	GSM1919117: mRNA11, Late log phase, Glucose carbon source; Bacteroides xylanisolvens; RNA-Seq
		SRR2827526.sra	GSM1919116: mRNA10, Late log phase, Glucose carbon source; Bacteroides xylanisolvens; RNA-Seq
		SRR2827527.sra	GSM1919115: mRNA6, Late log phase, Citrus Pectin carbon source; Bacteroides xylanisolvens; RNA-Seq
		SRR2827528.sra	GSM1919114: mRNA5, Late log phase, Citrus Pectin carbon source; Bacteroides xylanisolvens; RNA-Seq
		SRR2827529.sra	GSM1919113: mRNA4, Late log phase, Citrus Pectin carbon source; Bacteroides xylanisolvens; RNA-Seq
		SRR2827530.sra	GSM1919112: mRNA9, Late log phase, Apple Pectin carbon source; Bacteroides xylanisolvens; RNA-Seq
		SRR2827531.sra	GSM1919111: mRNA8, Late log phase, Apple Pectin carbon source; Bacteroides xylanisolvens; RNA-Seq
		SRR2827532.sra	GSM1919110: mRNA7, Late log phase, Apple Pectin carbon source; Bacteroides xylanisolvens; RNA-Seq
Escherichia coli	SRP069023	SRS1268205	GSM2049277: TW09308 transcriptome from starvation growth_2; Escherichia coli; RNA-Seq
		SRS1268206	GSM2049276: TW09308 transcriptome from batch growth_2; Escherichia coli; RNA-Seq
		SRS1268207	GSM2049275: TW09308 transcriptome from chemostat growth_2; Escherichia coli; RNA-Seq
		SRS1268208	GSM2049274: TW11588 transcriptome from starvation growth_2; Escherichia coli; RNA-Seq
		SRS1268209	GSM2049273: TW11588 transcriptome from batch growth_2; Escherichia coli; RNA-Seq
		SRS1268210	GSM2049272: TW11588 transcriptome from chemostat growth_2; Escherichia coli; RNA-Seq
		SRS1268176	GSM2049271: IAI1 transcriptome from starvation growth_2; Escherichia coli; RNA-Seq
		SRS1268177	GSM2049270: IAI1 transcriptome from batch growth_2; Escherichia coli; RNA-Seq
		SRS1268179	GSM2049269: IAI1 transcriptome from chemostat growth_2; Escherichia coli; RNA-Seq
		SRS1268185	GSM2049268: MG1655 transcriptome from starvation growth_2; Escherichia coli; RNA-Seq
		SRS1268178	GSM2049267: MG1655 transcriptome from batch growth_2; Escherichia coli; RNA-Seq
		SRS1268180	GSM2049266: MG1655 transcriptome from chemostat growth_2; Escherichia coli; RNA-Seq
		SRS1268181	GSM2049265: TW09308 transcriptome from starvation growth_1; Escherichia coli; RNA-Seq
		SRS1268182	GSM2049264: TW09308 transcriptome from batch growth_1; Escherichia coli; RNA-Seq
		SRS1268183	GSM2049263: TW09308 transcriptome from chemostat growth_1; Escherichia coli; RNA-Seq
		SRS1268184	GSM2049262: TW11588 transcriptome from starvation growth_1; Escherichia coli; RNA-Seq
		SRS1268175	GSM2049261: TW11588 transcriptome from batch growth_1; Escherichia coli; RNA-Seq
		SRS1268186	GSM2049260: TW11588 transcriptome from chemostat growth_1; Escherichia coli; RNA-Seq
		SRS1268187	GSM2049259: IAI1 transcriptome from starvation growth_1; Escherichia coli; RNA-Seq
		SRS1268188	GSM2049258: IAI1 transcriptome from batch growth_1; Escherichia coli; RNA-Seq
		SRS1268189	GSM2049257: IAI1 transcriptome from chemostat growth_1; Escherichia coli; RNA-Seq
		SRS1268191	GSM2049256: MG1655 transcriptome from starvation growth_1; Escherichia coli; RNA-Seq
		SRS1268190	GSM2049255: MG1655 transcriptome from batch growth_1; Escherichia coli; RNA-Seq

		SRS1268192	GSM2049254: MG1655 transcriptome from chemostat growth_1; Escherichia coli; RNA-Seq
	SRP056663	SRS886944	GSM1646318: Glucose time course, 336 hour time point, biological replicate 3, rRNA not depleted; Escherichia coli; RNA-Seq
		SRS886945	GSM1646317: Glucose time course, 168 hour time point, biological replicate 3, rRNA not depleted; Escherichia coli; RNA-Seq
		SRS886946	GSM1646316: Glucose time course, 48 hour time point, biological replicate 3, rRNA not depleted; Escherichia coli; RNA-Seq
		SRS886947	GSM1646315: Glucose time course, 24 hour time point, biological replicate 3, rRNA not depleted; Escherichia coli; RNA-Seq
		SRS886948	GSM1646314: Glucose time course, 8 hourt ime point, biological replicate 3, rRNA not depleted; Escherichia coli; RNA-Seq
		SRS886949	GSM1646313: Glucose time course, 6 hour time point, biological replicate 3, rRNA not depleted; Escherichia coli; RNA-Seq
		SRS886950	GSM1646312: Glucose time course, 5 hour time point, biological replicate 3, rRNA not depleted; Escherichia coli; RNA-Seq
		SRS886951	GSM1646311: Glucose time course, 4 hour time point, biological replicate 3, rRNA not depleted; Escherichia coli; RNA-Seq
		SRS886952	GSM1646310: Glucose time course, 3 hour time point, biological replicate 3, rRNA not depleted; Escherichia coli; RNA-Seq
		SRS886959	GSM1646309: Glucose time course, 336 hour time point, biological replicate 3; Escherichia coli; RNA-Seq
		SRS886953	GSM1646308: Glucose time course, 168 hour time point, biological replicate 3; Escherichia coli; RNA-Seq
		SRS886955	GSM1646307: Glucose time course, 48 hour time point, biological replicate 3; Escherichia coli; RNA-Seq
		SRS886954	GSM1646306: Glucose time course, 24 hour time point, biological replicate 3; Escherichia coli; RNA-Seq
		SRS886956	GSM1646305: Glucose time course, 8 hourt ime point, biological replicate 3; Escherichia coli; RNA-Seq
		SRS886957	GSM1646304: Glucose time course, 6 hour time point, biological replicate 3; Escherichia coli; RNA-Seq
		SRS886958	GSM1646303: Glucose time course, 5 hour time point, biological replicate 3; Escherichia coli; RNA-Seq
		SRS886960	GSM1646302: Glucose time course, 4 hour time point, biological replicate 3; Escherichia coli; RNA-Seq
		SRS886961	GSM1646301: Glucose time course, 3 hour time point, biological replicate 3; Escherichia coli; RNA-Seq
		SRS886962	GSM1646300: Glucose time course, 336 hour time point, biological replicate 2; Escherichia coli; RNA-Seq
		SRS886963	GSM1646299: Glucose time course, 168 hour time point, biological replicate 2; Escherichia coli; RNA-Seq
		SRS886964	GSM1646298: Glucose time course, 48 hour time point, biological replicate 2; Escherichia coli; RNA-Seq
		SRS886965	GSM1646297: Glucose time course, 24 hour time point, biological replicate 2; Escherichia coli; RNA-Seq
		SRS886968	GSM1646296: Glucose time course, 8 hour time point, biological replicate 2; Escherichia coli; RNA-Seq
		SRS886967	GSM1646295: Glucose time course, 6 hour time point, biological replicate 2; Escherichia coli; RNA-Seq
		SRS886966	GSM1646294: Glucose time course, 5 hour time point, biological replicate 2; Escherichia coli; RNA-Seq
		SRS886970	GSM1646293: Glucose time course, 4 hour time point, biological replicate 2; Escherichia coli; RNA-Seq
		SRS886969	GSM1646292: Glucose time course, 3 hour time point, biological replicate 2; Escherichia coli; RNA-Seq
		SRS886971	GSM1646291: Glucose time course, 336 hour time point, biological replicate 1; Escherichia coli; RNA-Seq
		SRS886974	GSM1646290: Glucose time course, 168 hour time point, biological replicate 1; Escherichia coli; RNA-Seq
		SRS886972	GSM1646289: Glucose time course, 48 hour time point, biological replicate 1; Escherichia coli; RNA-Seq
		SRS886973	GSM1646288: Glucose time course, 24 hour time point, biological replicate 1; Escherichia coli; RNA-Seq
		SRS886976	GSM1646287: Glucose time course, 8 hour time point, biological replicate 1; Escherichia coli; RNA-Seq
		SRS886975	GSM1646286: Glucose time course, 6 hour time point, biological replicate 1; Escherichia coli; RNA-Seq
		SRS886977	GSM1646285: Glucose time course, 5 hour time point, biological replicate 1; Escherichia coli; RNA-Seq
		SRS886979	GSM1646284: Glucose time course, 4 hour time point, biological replicate 1; Escherichia coli; RNA-Seq
		SRS886978	GSM1646283: Glucose time course, 3 hour time point, biological replicate 1; Escherichia coli; RNA-Seq
	SRP043192	SRS634267	Exponential growing DS1 Strain - Biological replicate
		SRS634267	Exponential growing DS1 Strain.
		SRS634267	Type II persister cells of DS1 Strain - Biological replicate.
		SRS634267	Type II persister cells of DS1 Strain.
		SRS639031	GSM1413881: rpoS_N_strv_TEX; Escherichia coli BW38028; RNA-Seq
		SRS639030	GSM1413880: WT_N_strv_TEX; Escherichia coli BW38028; RNA-Seq
		SRS639029	GSM1413879: WT_glucose_stat_TEX; Escherichia coli BW38028; RNA-Seq
		SRS639028	GSM1413878: WT_glucose_log_TEX; Escherichia coli BW38028; RNA-Seq
		SRS639027	GSM1413877: rpoS_N_strv; Escherichia coli BW38028; RNA-Seq
		SRS639026	GSM1413876: WT_N_strv; Escherichia coli BW38028; RNA-Seq
		SRS639025	GSM1413875: WT_glucose_stat; Escherichia coli BW38028; RNA-Seq
		SRS639024	GSM1413874: WT_glucose_log; Escherichia coli BW38028; RNA-Seq
Mycobacterium tuberculosis	SRP056290	SRS875615	GSM1636561: HN878 RIF-R mutant with RpoB:S531L, replicate 3; Mycobacterium tuberculosis HN878; RNA-Seq
		SRS875616	GSM1636560: HN878 RIF-R mutant with RpoB:S531L, replicate 2; Mycobacterium tuberculosis HN878; RNA-Seq
		SRS875617	GSM1636559: HN878 RIF-R mutant with RpoB:S531L, replicate 1; Mycobacterium tuberculosis HN878; RNA-Seq
		SRS875619	GSM1636558: HN878 RIF-R mutant with RpoB:H526R, replicate 3; Mycobacterium tuberculosis HN878; RNA-Seq
		SRS875618	GSM1636557: HN878 RIF-R mutant with RpoB:H526R, replicate 2; Mycobacterium tuberculosis HN878; RNA-Seq
		SRS875620	GSM1636556: HN878 RIF-R mutant with RpoB:H526R, replicate 1; Mycobacterium tuberculosis HN878; RNA-Seq
		SRS875621	GSM1636555: HN878 RIF-R mutant with RpoB:H526D, replicate 1; Mycobacterium tuberculosis HN878; RNA-Seq

		SRS875622	GSM1636554: HN878 RIF-R mutant with RpoB:D516V, replicate 3; Mycobacterium tuberculosis HN878; RNA-Seq
		SRS875623	GSM1636553: HN878 RIF-R mutant with RpoB:D516V, replicate 2; Mycobacterium tuberculosis HN878; RNA-Seq
		SRS875624	GSM1636552: HN878 RIF-R mutant with RpoB:D516V, replicate 1; Mycobacterium tuberculosis HN878; RNA-Seq
		SRS875632	GSM1636551: HN878 Beijing strain, replicate 3; Mycobacterium tuberculosis HN878; RNA-Seq
		SRS875625	GSM1636550: HN878 Beijing strain, replicate 2; Mycobacterium tuberculosis HN878; RNA-Seq
		SRS875626	GSM1636549: HN878 Beijing strain, replicate 1; Mycobacterium tuberculosis HN878; RNA-Seq
		SRS875629	GSM1636548: H37Rv grown in toloxapol, pH 5.5, replicate 2; Mycobacterium tuberculosis H37Rv; RNA-Seq
		SRS875627	GSM1636547: H37Rv grown in toloxapol, pH 5.5, replicate 1; Mycobacterium tuberculosis H37Rv; RNA-Seq
		SRS875631	GSM1636546: H37Rv grown in toloxapol, pH 7.0, replicate 2; Mycobacterium tuberculosis H37Rv; RNA-Seq
		SRS875633	GSM1636545: H37Rv grown in toloxapol, pH 7.0, replicate 1; Mycobacterium tuberculosis H37Rv; RNA-Seq
		SRS875630	GSM1636544: H37Rv grown in low iron, 1 week replicate 3; Mycobacterium tuberculosis H37Rv; RNA-Seq
		SRS875628	GSM1636543: H37Rv grown in low iron, 1 week replicate 2; Mycobacterium tuberculosis H37Rv; RNA-Seq
		SRS875634	GSM1636542: H37Rv grown in low iron, 1 week replicate 1; Mycobacterium tuberculosis H37Rv; RNA-Seq
		SRS875635	GSM1636541: H37Rv grown in low iron, 1 day, replicate 3; Mycobacterium tuberculosis H37Rv; RNA-Seq
		SRS875637	GSM1636540: H37Rv grown in low iron, 1 day, replicate 2; Mycobacterium tuberculosis H37Rv; RNA-Seq
		SRS875636	GSM1636539: H37Rv grown in low iron, 1 day, replicate 1; Mycobacterium tuberculosis H37Rv; RNA-Seq
		SRS875638	GSM1636538: H37Rv grown in high iron, replicate 3; Mycobacterium tuberculosis H37Rv; RNA-Seq
		SRS875640	GSM1636537: H37Rv grown in high iron, replicate 2; Mycobacterium tuberculosis H37Rv; RNA-Seq
		SRS875639	GSM1636536: H37Rv grown in high iron, replicate 1; Mycobacterium tuberculosis H37Rv; RNA-Seq
		SRS875645	GSM1636535: H37Rv grown on 0.4% glucose, replicate 3; Mycobacterium tuberculosis H37Rv; RNA-Seq
		SRS875641	GSM1636534: H37Rv grown on 0.4% glucose, replicate 2; Mycobacterium tuberculosis H37Rv; RNA-Seq
		SRS875614	GSM1636533: H37Rv grown on 0.4% glucose, replicate 1; Mycobacterium tuberculosis H37Rv; RNA-Seq
		SRS875613	GSM1636532: H37Rv grown on 0.2% glucose+0.1% butyrate, replicate 3; Mycobacterium tuberculosis H37Rv; RNA-Seq
		SRS875644	GSM1636531: H37Rv grown on 0.2% glucose+0.1% butyrate, replicate 2; Mycobacterium tuberculosis H37Rv; RNA-Seq
		SRS875643	GSM1636530: H37Rv grown on 0.2% glucose+0.1% butyrate, replicate 1; Mycobacterium tuberculosis H37Rv; RNA-Seq
		SRS875642	GSM1636529: H37Rv grown on 0.1% butyrate, replicate 3; Mycobacterium tuberculosis H37Rv; RNA-Seq
		SRS875646	GSM1636528: H37Rv grown on 0.1% butyrate, replicate 2; Mycobacterium tuberculosis H37Rv; RNA-Seq
		SRS875612	GSM1636527: H37Rv grown on 0.1% butyrate, replicate 1; Mycobacterium tuberculosis H37Rv; RNA-Seq
	SRP032513	SRS498419	GSM1257648: Pyruvate pH 5.7, Rep B; Mycobacterium tuberculosis; RNA-Seq
		SRS498418	GSM1257647: Pyruvate pH 5.7, Rep A; Mycobacterium tuberculosis; RNA-Seq
		SRS498417	GSM1257646: Pyruvate pH 7.0, Rep B; Mycobacterium tuberculosis; RNA-Seq
		SRS498416	GSM1257645: Pyruvate pH 7.0, Rep A; Mycobacterium tuberculosis; RNA-Seq
		SRS498415	GSM1257644: Glycerol pH 5.7, Rep B; Mycobacterium tuberculosis; RNA-Seq
		SRS498414	GSM1257643: Glycerol pH 5.7, Rep A; Mycobacterium tuberculosis; RNA-Seq
		SRS498413	GSM1257642: Glycerol pH 7.0, Rep B; Mycobacterium tuberculosis; RNA-Seq
		SRS498412	GSM1257641: Glycerol pH 7.0, Rep A; Mycobacterium tuberculosis; RNA-Seq
	SRP056155	SRS874204	GSM1633741: Resuscitation rep3; Mycobacterium bovis BCG str. Pasteur 1173P2; RNA-Seq
		SRS874205	GSM1633740: Resuscitation rep2; Mycobacterium bovis BCG str. Pasteur 1173P2; RNA-Seq
		SRS874206	GSM1633739: Resuscitation rep1; Mycobacterium bovis BCG str. Pasteur 1173P2; RNA-Seq
		SRS874207	GSM1633738: Starvation day20 rep3; Mycobacterium bovis BCG str. Pasteur 1173P2; RNA-Seq
		SRS874208	GSM1633737: Starvation day20 rep2; Mycobacterium bovis BCG str. Pasteur 1173P2; RNA-Seq
		SRS874209	GSM1633736: Starvation day20 rep1; Mycobacterium bovis BCG str. Pasteur 1173P2; RNA-Seq
		SRS874210	GSM1633735: Starvation day10 rep3; Mycobacterium bovis BCG str. Pasteur 1173P2; RNA-Seq
		SRS874211	GSM1633734: Starvation day10 rep2; Mycobacterium bovis BCG str. Pasteur 1173P2; RNA-Seq
		SRS874212	GSM1633733: Starvation day10 rep1; Mycobacterium bovis BCG str. Pasteur 1173P2; RNA-Seq
		SRS874213	GSM1633732: Starvation day4 rep3; Mycobacterium bovis BCG str. Pasteur 1173P2; RNA-Seq
		SRS874214	GSM1633731: Starvation day4 rep2; Mycobacterium bovis BCG str. Pasteur 1173P2; RNA-Seq
		SRS874215	GSM1633730: Starvation day4 rep1; Mycobacterium bovis BCG str. Pasteur 1173P2; RNA-Seq
		SRS874216	GSM1633729: Log phase control rep3; Mycobacterium bovis BCG str. Pasteur 1173P2; RNA-Seq
		SRS874217	GSM1633728: Log phase control rep2; Mycobacterium bovis BCG str. Pasteur 1173P2; RNA-Seq
		SRS874218	GSM1633727: Log phase control rep1; Mycobacterium bovis BCG str. Pasteur 1173P2; RNA-Seq

15 - Table 8.9 – Retrieved RNAseq datasets

A list of the different RNAseq datasets that were retrieved in this study, listing the organism, the project and the individual sample ids. The description of the sample form the producer of the data is included.

Appendix

Differentially expressed proteins – Split Cluster vs Wildtype (6AA)					
Protein Name	Protein Symbol	Log2 WT Average Counts	Log2 Split Cluster Average Counts	Difference	-Log2 P-value
2-iminobutanoate/2-iminopropanoate deaminase	yabJ	4.60	0.67	3.94	3.32
Catalase-2	katE	3.09	0.67	2.42	2.49
General stress protein 69	yhdN	2.97	0.67	2.31	1.31
Oxalate decarboxylase OxdC	oxdC	3.66	1.53	2.13	2.64
Uncharacterized N-acetyltransferase YvbK	yvbK	2.11	0.00	2.11	4.40
UPF0331 protein YutE	yutE	1.97	0.00	1.97	3.12
FMN-dependent NADPH-azoreductase	azr	2.25	0.33	1.92	1.82
3-dehydroquinate dehydratase	aroD	2.57	0.67	1.90	1.31
tRNA N6-adenosine threonylcarbamoyltransferase	tsaD	2.11	0.33	1.77	2.15
Primosomal protein DnaI	dnaI	2.41	0.67	1.74	1.22
Uncharacterized protein YydD	yydD	3.69	1.97	1.72	2.71
Uncharacterized protein YddK	yddK	3.11	1.39	1.72	2.98
Teichoic acid translocation permease protein TagG	tagG	1.72	0.00	1.72	3.62
Ribosomal protein L11 methyltransferase	prmA	2.24	0.53	1.71	1.26
Flagellar hook-associated protein 2	fliD	3.00	1.33	1.67	2.13
SPBc2 prophage-derived uncharacterized protein YopC	yopC	2.00	0.33	1.67	2.13
Guanine/hypoxanthine permease PbuG	pbuG	2.00	0.33	1.67	2.13
SPBc2 prophage-derived uncharacterized protein YonI	yonI	1.67	0.00	1.67	2.13
Putative niacin/nicotinamide transporter NaiP	naiP	1.67	0.00	1.67	2.13
Putative phosphinothricin acetyltransferase YwnH	ywnH	2.50	0.86	1.64	1.59
General stress protein 17M	yflT	3.04	1.44	1.60	1.58
UPF0111 protein YkaA	ykaA	3.41	1.86	1.55	3.16
General stress protein CTC	ctc	3.53	1.99	1.54	1.63
Redox-sensing transcriptional repressor Rex	rex	1.86	0.33	1.53	1.88
Regulatory protein RecX	recX	1.86	0.33	1.53	1.88
Putative metal-dependent hydrolase YfiT	yfiT	2.30	0.86	1.44	1.37
Putative ribonuclease YwqJ	ywqJ	2.11	0.67	1.44	1.83
Phosphoribosylglycinamide formyltransferase	purN	4.19	2.77	1.42	2.24
CCA-adding enzyme	cca	2.08	0.67	1.41	1.57
Antitoxin YxxD	yxxD	2.27	0.86	1.41	1.23
UPF0755 protein YrrL	yrrL	3.32	1.92	1.40	1.82
Protoporphyrinogen oxidase	hemY	1.39	0.00	1.39	2.69
UPF0118 membrane protein YueF	yueF	1.39	0.00	1.39	2.69
Uncharacterized protein YqkB	yqkB	1.39	0.00	1.39	2.69
Uncharacterized oxidoreductase YcsN	ycsN	1.39	0.00	1.39	2.69
sp|P94498|CYSH1_BACSU Phosphoadenosine phosphosulfate reductase	cysH	3.70	2.32	1.38	4.55
RNA polymerase sigma-B factor	sigB	3.20	1.83	1.37	1.91
SPBc2 prophage-derived uncharacterized protein YopQ	yopQ	3.36	2.00	1.36	3.84
Gluconokinase	gntK	2.21	0.86	1.35	1.33
DNA-binding protein HU 1	hupA	6.78	5.44	1.34	4.30
HTH-type transcriptional regulator YodB	yodB	1.33	0.00	1.33	1.79
Hypoxanthine-guanine phosphoribosyltransferase	hprT	5.05	3.72	1.33	2.96
Anti-sigma-B factor antagonist	rsbV	2.71	1.39	1.32	2.03
Dihydroorotate dehydrogenase B (NAD(+)), electron transfer subunit	pyrK	3.17	1.86	1.31	3.16
Uncharacterized protein YvbH	yvbH	2.50	1.19	1.30	2.43
Flagellar protein FliL	fliL	1.97	0.67	1.30	1.52
Uncharacterized protein YvfG	yvfG	3.06	1.83	1.23	2.08
5'-3' exonuclease	ypcP	2.21	1.00	1.21	3.46
Purine nucleoside phosphorylase 1	punA	4.26	3.06	1.20	2.91
Arginine repressor	argR	2.39	1.19	1.19	1.91
Putative lipoprotein YvcA	yvcA	1.86	0.67	1.19	1.53
Glucose 1-dehydrogenase 2	ycdF	1.53	0.33	1.19	1.27
Citrate synthase 1	citA	1.19	0.00	1.19	2.44
High-affinity proline transporter PutP	putP	1.19	0.00	1.19	2.44
Probable glucosamine-6-phosphate deaminase 2	gamA	1.19	0.00	1.19	2.44
Uncharacterized PIN and TRAM-domain containing protein YacL	yacL	1.19	0.00	1.19	2.44
Putative exported peptide YydF	yydF	1.19	0.00	1.19	2.44
Central glycolytic genes regulator	cggR	3.58	2.39	1.19	1.31
Type-2 restriction enzyme BsuMI component YdiS	ydiS	2.57	1.39	1.18	2.10
General stress protein 20U	dps	3.94	2.77	1.18	1.71
Uncharacterized protein YtoQ	ytoQ	3.03	1.86	1.16	1.94
Uncharacterized ABC transporter ATP-binding protein YhaQ	yhaQ	1.83	0.67	1.16	1.32
Oligopeptide transport system permease protein OppC	oppC	1.83	0.67	1.16	1.32

Xylose isomerase	xylA	4.60	3.44	1.16	2.73
DNA-directed RNA polymerase subunit delta	rpoE	3.72	2.57	1.15	2.39
Thioredoxin-like protein YdbP	ydbP	3.66	2.53	1.13	1.75
Uncharacterized protein YjlC	yjlC	6.12	5.00	1.12	3.55
Protein translocase subunit SecY	secY	2.65	1.53	1.12	1.55
Putative methyl-accepting chemotaxis protein YoaH	yoaH	2.66	1.58	1.07	3.88
Probable anti-sigma-M factor YhdL	yhdL	3.04	1.97	1.07	1.83
Multidrug resistance protein 3	bmr3	1.72	0.67	1.06	1.37
Uncharacterized membrane protein YubF	yubF	1.39	0.33	1.06	1.28
Probable enoyl-CoA hydratase	fadB	1.39	0.33	1.06	1.28
Minor teichoic acid biosynthesis protein GgaA	ggaA	1.39	0.33	1.06	1.28
Quinol oxidase subunit 3	qoxC	1.39	0.33	1.06	1.28
Uncharacterized protein YukE	yukE	5.56	4.55	1.02	2.44
Exodeoxyribonuclease 7 small subunit	xseB	3.58	2.57	1.01	2.53
Flagellar motor switch protein FliM	fliM	2.73	1.72	1.01	2.52
Succinate-semialdehyde dehydrogenase [NADP(+)]	gabD	2.87	1.86	1.01	2.57
sp|O06478|ALDH5_BACSU Putative aldehyde dehydrogenase YfmT	yfmT	4.95	3.96	1.00	2.69
Pyridoxine kinase	pdxK	4.64	3.66	0.98	3.22
Uncharacterized protein YcsD	ycsD	1.97	1.00	0.97	1.98
3-isopropylmalate dehydratase small subunit	leuD	4.33	3.37	0.96	2.30
sp|O32243|OPUCC_BACSU Glycine betaine/carnitine/choline-binding protein OpuCC	opuCC	3.50	2.55	0.95	1.84
3-dehydroquinate synthase	aroB	5.34	4.39	0.95	3.19
Homoserine kinase	thrB	4.60	3.66	0.94	2.25
Octanoyltransferase LipM	lipM	3.05	2.11	0.94	2.36
Putative O-methyltransferase YrrM	yrrM	1.58	0.67	0.92	1.29
Aminomethyltransferase	gcvT	2.11	1.19	0.91	1.83
Xanthine permease	pbuX	2.11	1.19	0.91	1.83
Putative sensory transducer protein YfmS	yfmS	4.32	3.41	0.91	3.14
Cysteine--tRNA ligase	cysS	4.41	3.50	0.91	2.65
50S ribosomal protein L29	rpmC	5.06	4.16	0.90	1.91
Gamma-glutamyl phosphate reductase	proA	4.29	3.41	0.88	3.41
Probable 2-ketogluconate reductase	yvcT	2.97	2.11	0.87	1.76
Stage 0 sporulation protein A	spo0A	1.86	1.00	0.86	2.47
Uncharacterized HTH-type transcriptional regulator YvdT	yvdT	1.86	1.00	0.86	2.47
Ribonuclease 3	rnc	2.71	1.86	0.85	1.59
UPF0234 protein yitk	yitK	3.84	2.99	0.84	2.46
Uncharacterized protein YjoA	yjoA	3.25	2.41	0.84	1.91
DNA topoisomerase 3	topB	4.05	3.21	0.84	1.91
Putative heme-dependent peroxidase YwfI	ywfI	2.80	1.97	0.83	1.56
SPBc2 prophage-derived uncharacterized protein YokE	yokE	2.21	1.39	0.82	1.68
Uncharacterized protein YfjT	yfjT	2.21	1.39	0.82	1.68
Putative acetyltransferase YjbC	yjbC	2.19	1.39	0.81	1.36
Putative HMP/thiamine permease protein YkoC	ykoC	2.19	1.39	0.81	1.36
Uncharacterized protein YqeK	yqeK	2.19	1.39	0.81	1.36
Uncharacterized protein YvyC	yvyC	2.00	1.19	0.81	1.84
Glutamine synthetase	glnA	7.32	6.53	0.80	2.41
DNA translocase SftA	sftA	3.84	3.04	0.80	2.18
50S ribosomal protein L16	rplP	6.45	5.65	0.80	4.97
Serine hydroxymethyltransferase	glyA	6.52	5.73	0.79	2.67
Putative cysteine ligase BshC	bshC	3.80	3.01	0.79	1.42
GTP pyrophosphokinase	relA	4.71	3.93	0.78	2.62
Putative cysteine protease YraA	yraA	3.93	3.16	0.77	2.47
Chemotaxis protein CheV	cheV	4.95	4.20	0.76	4.46
Ftsk domain-containing protein YukB	yukB	2.97	2.21	0.75	1.47
Carboxypeptidase 1	ypwA	4.30	3.54	0.75	4.00
Chromosome partition protein Smc	smc	3.37	2.64	0.73	1.65
Pyridoxal 5'-phosphate synthase subunit PdxT	pdxT	4.03	3.31	0.72	2.01
sp|P39576|ILVE2_BACSU Branched-chain-amino-acid aminotransferase 2	ilvK	5.80	5.08	0.72	2.62
UDP-N-acetylenolpyruvoylglucosamine reductase	murB	4.34	3.62	0.72	3.23
Putative N-acetylmuramoyl-L-alanine amidase YrvJ	yrvJ	2.11	1.39	0.72	1.49
UTP--glucose-1-phosphate uridylyltransferase	gtaB	5.69	4.98	0.71	2.43
GTP-sensing transcriptional pleiotropic repressor CodY	codY	5.10	4.39	0.71	3.13
Serine-protein kinase RsbW	rsbW	3.77	3.06	0.71	2.52
HTH-type transcriptional regulator Hpr	hpr	2.57	1.86	0.71	1.64
sp|O34577|CYSC1_BACSU Probable adenylyl-sulfate kinase	cysC	3.20	2.50	0.70	1.49
Glyoxal reductase	yvgN	4.96	4.27	0.70	2.22
Methionine--tRNA ligase	metG	5.81	5.11	0.69	2.61

sp|O34916|DAPEL_BACSU N-acetyldiaminopimelate deacetylase	ykuR	2.41	1.72	0.69	1.86
Manganese transport system ATP-binding protein MntB	mntB	5.09	4.41	0.67	3.90
sp|P39120|CISY2_BACSU Citrate synthase 2	citZ	4.39	3.72	0.66	1.91
Uncharacterized protein YpuA	ypuA	3.32	2.66	0.66	2.38
2,3,4,5-tetrahydropyridine-2,6-dicarboxylate N-acetyltransferase	dapH	5.26	4.60	0.66	3.16
30S ribosomal protein S6	rpsF	6.67	6.01	0.66	2.15
Flagellar motor switch protein FliG	fliG	3.15	2.50	0.65	1.56
Proline--tRNA ligase	proS	5.70	5.05	0.65	2.43
Putative transcriptional regulator YwtF	ywtF	3.05	2.41	0.64	1.89
Arginase	rocF	6.91	6.27	0.64	2.33
Ribonuclease J1	rnjA	6.01	5.38	0.64	3.35
sp|P39816|PTW3C_BACSU Putative PTS system glucosamine-specific EIICBA component	gamP	2.50	1.86	0.64	1.75
Metalloregulation DNA-binding stress protein	mrgA	2.94	2.30	0.63	1.60
L-lactate dehydrogenase	ldh	4.66	4.03	0.63	2.46
Glucose-6-phosphate 1-dehydrogenase	zwf	5.32	4.69	0.63	1.87
Threonine--tRNA ligase 1	thrS	5.89	5.27	0.62	2.15
DEAD-box ATP-dependent RNA helicase CshB	cshB	4.68	4.06	0.62	1.52
Sirohydrochlorin ferrochelatase	sirB	2.94	2.32	0.61	3.17
Zinc-transporting ATPase	zosA	4.56	3.96	0.60	1.63
sp|P70970|ECFA2_BACSU Energy-coupling factor transporter ATP-binding protein EcfA2	ecfAB	2.32	1.72	0.60	1.91
Methyl-accepting chemotaxis protein TlpB	tlpB	4.56	3.96	0.60	2.14
Methyl-accepting chemotaxis protein McpB	mcpB	5.46	4.87	0.59	2.12
Alkyl hydroperoxide reductase subunit C	ahpC	8.02	7.43	0.59	2.52
Asparagine synthetase [glutamine-hydrolyzing] 1	asnB	5.30	4.71	0.59	1.77
NADPH-dependent 7-cyano-7-deazaguanine reductase	queF	3.50	2.92	0.58	1.51
Uncharacterized membrane protein YrrS	yrrS	2.99	2.41	0.58	1.89
Uncharacterized zinc protease YmfH	ymfH	3.22	2.65	0.58	1.56
Putative carbonic anhydrase YtiB	ytiB	3.50	2.93	0.57	1.96
Dihydroorotase	pyrC	4.14	3.57	0.57	1.79
RNA polymerase sigma factor SigA	sigA	5.04	4.48	0.57	3.10
Uncharacterized protein ymdB	ymdB	3.50	2.94	0.56	2.18
1-acyl-sn-glycerol-3-phosphate acyltransferase	plsC	3.41	2.86	0.55	1.71
Ribosome-binding factor A	rbfA	2.66	2.11	0.55	1.87
Chemotaxis protein CheA	cheA	5.09	4.54	0.55	1.91
Flagellar motor switch phosphatase FliY	fliY	4.60	4.05	0.55	1.84
Adenylate kinase	adk	5.46	4.91	0.55	1.94
Enolase	eno	8.23	7.69	0.54	2.29
Tryptophan--tRNA ligase	trpS	3.41	2.87	0.54	2.11
UDP-glucose 4-epimerase	galE	3.90	3.37	0.54	2.76
Valine--tRNA ligase	valS	6.19	5.65	0.54	2.48
UPF0435 protein YfkK	yfkK	3.54	3.00	0.54	2.06
L-cystine-binding protein TcyA	tcyA	4.58	4.05	0.53	2.25
2-oxoglutarate dehydrogenase E1 component	odhA	5.80	5.27	0.53	2.66
FeS cluster assembly protein SufD	sufD	6.76	6.23	0.53	2.48
Protein YtsP	ytsP	2.94	2.41	0.53	2.08
Adenylosuccinate lyase	purB	6.31	5.79	0.52	3.15
Superoxide dismutase [Mn]	sodA	5.44	4.92	0.52	2.68
Tyrosine--tRNA ligase 1	tyrS1	5.55	5.04	0.51	3.16
2-isopropylmalate synthase	leuA	7.07	6.56	0.51	3.33
tRNA modification GTPase MnmE	mnmE	3.62	3.11	0.50	1.91
Glucose-6-phosphate isomerase	pgi	6.54	6.04	0.50	2.45
Isocitrate dehydrogenase [NADP]	icd	6.36	5.86	0.50	3.85
Cluster of Methyl-accepting chemotaxis protein McpA	mcpA	5.18	4.68	0.50	1.90
Orotidine 5'-phosphate decarboxylase	pyrF	4.64	4.14	0.49	2.05
4-hydroxy-tetrahydrodipicolinate synthase	dapA	5.40	4.90	0.49	1.83
Catabolite control protein A	ccpA	5.12	4.62	0.49	3.10
Phosphoenolpyruvate-protein phosphotransferase	ptsI	7.19	6.70	0.49	2.48
Lactate utilization protein A	lutA	3.90	3.41	0.49	2.62
Aspartate-semialdehyde dehydrogenase	asd	6.10	5.61	0.49	2.56
GMP synthase [glutamine-hydrolyzing]	guaA	6.28	5.79	0.49	1.70
Glutamyl-tRNA(Gln) amidotransferase subunit A	gatA	6.18	5.70	0.48	2.07
Uncharacterized protein YxkC	yxkC	6.40	5.93	0.47	2.01
Acetolactate synthase large subunit	ilvB	6.31	5.85	0.45	2.25
ATP-dependent zinc metalloprotease FtsH	ftsH	6.12	5.67	0.45	1.78
Oligopeptide transport ATP-binding protein OppF	oppF	3.66	3.22	0.44	2.09
Exodeoxyribonuclease 7 large subunit	xseA	2.94	2.50	0.44	1.80

sp|P94390|PROD2_BACSU Proline dehydrogenase 2	putB	4.54	4.11	0.43	2.05
Transaldolase	tal	5.93	5.51	0.42	1.86
Glycerol-3-phosphate dehydrogenase [NAD(P)+]	gpsA	3.87	3.46	0.42	2.14
sp|O34992|OPUCA_BACSU Glycine betaine/carnitine/choline transport ATP-binding protein OpuCA	opuCA	5.18	4.77	0.42	1.86
Negative regulator of genetic competence ClpC/MecB	clpC	6.12	5.71	0.41	2.33
Putative nitrogen fixation protein YutI	yutI	2.41	2.00	0.41	2.02
sp|Q01625|MISCA_BACSU Membrane protein insertase MisCA	misCA	2.41	2.00	0.41	2.02
Phage shock protein A homolog	ydjF	4.44	4.03	0.41	2.82
Leucine--tRNA ligase	leuS	5.35	4.95	0.39	2.67
30S ribosomal protein S17	rpsQ	5.85	5.47	0.38	2.06
ATP-dependent Clp protease ATP-binding subunit ClpX	clpX	5.51	5.14	0.37	2.62
Pyruvate carboxylase	pyc	7.04	6.67	0.37	3.23
DNA-directed RNA polymerase subunit beta'	rpoC	7.48	7.11	0.37	4.12
30S ribosomal protein S4	rpsD	7.01	6.66	0.35	2.21
ATP synthase subunit beta	atpD	6.65	6.30	0.35	3.46
ATP-dependent Clp protease proteolytic subunit	clpP	4.72	4.37	0.35	2.55
Elongation factor Ts	tsf	7.35	7.01	0.34	3.99
Elongation factor G	fusA	8.58	8.24	0.34	3.02
Phenylalanine--tRNA ligase beta subunit	pheT	6.82	6.48	0.33	3.76
Sulfite reductase [NADPH] flavoprotein alpha-component	cysJ	4.60	4.27	0.33	2.45
Alanine--tRNA ligase	alaS	5.56	5.24	0.33	2.69
Uncharacterized ABC transporter solute-binding protein YclQ	yclQ	4.74	4.44	0.30	3.32
50S ribosomal protein L5	rplE	7.09	6.80	0.29	2.75
Chorismate synthase	aroC	5.58	5.89	-0.31	3.43
Dihydrolipoyllysine-residue acetyltransferase component of pyruvate dehydrogenase complex	pdhC	7.36	7.70	-0.34	2.35
Transketolase	tkt	6.68	7.04	-0.37	2.75
50S ribosomal protein L19	rplS	5.95	6.33	-0.37	2.35
tRNA nuclease WapA	wapA	5.95	6.34	-0.39	2.32
50S ribosomal protein L30	rpmD	4.11	4.52	-0.41	2.42
50S ribosomal protein L20	rplT	4.29	4.74	-0.45	1.84
Phosphoglucosamine mutase	glmM	4.32	4.79	-0.47	1.96
Succinyl-CoA ligase [ADP-forming] subunit alpha	sucD	4.48	4.96	-0.48	1.82
Putative phosphatase YitU	yitU	2.50	3.00	-0.50	2.34
Immunity protein WapI	wapI	2.99	3.50	-0.50	1.68
UPF0173 metal-dependent hydrolase YtkL	ytkL	3.22	3.73	-0.51	2.05
Cysteine desulfurase SufS	sufS	3.96	4.50	-0.53	2.05
Putative ABC transporter substrate-binding lipoprotein YhfQ	yhfQ	2.72	3.27	-0.55	1.69
Thiazole tautomerase	tenI	2.94	3.49	-0.56	1.92
Malonyl CoA-acyl carrier protein transacylase	fabD	3.70	4.26	-0.56	1.83
Acireductone dioxygenase	mtnD	2.94	3.50	-0.56	2.18
Putative dipeptidase YkvY	ykvY	3.05	3.62	-0.57	1.98
Uncharacterized protein YcnI	ycnI	2.21	2.81	-0.59	2.28
Uncharacterized protein YwnB	ywnB	1.72	2.32	-0.60	1.91
Fumarate hydratase class II	fumC	4.41	5.03	-0.61	2.56
UDP-N-acetylglucosamine 2-epimerase	mnaA	2.48	3.11	-0.62	1.49
Ribose-phosphate pyrophosphokinase	prs	4.77	5.41	-0.64	2.01
50S ribosomal protein L4	rplD	5.85	6.52	-0.68	2.87
sp|Q08788|SRFAD_BACSU Surfactin synthase thioesterase subunit	srfAD	1.72	2.41	-0.69	1.86
Aspartokinase 2	lysC	4.78	5.48	-0.70	1.63
30S ribosomal protein S8	rpsH	4.97	5.69	-0.72	2.05
Uncharacterized protein YneT	yneT	1.39	2.11	-0.72	1.49
Uncharacterized protease YrrO	yrrO	2.21	2.94	-0.72	2.35
4-hydroxy-tetrahydrodipicolinate reductase	dapB	3.04	3.77	-0.72	1.82
Putative NAD(P)H nitroreductase YodC	yodC	3.20	3.93	-0.73	1.47
Putative metal chaperone YciC	yciC	6.75	7.49	-0.73	2.94
Methionyl-tRNA formyltransferase	fmt	4.18	4.92	-0.74	2.22
30S ribosomal protein S10	rpsJ	5.01	5.76	-0.75	3.37
50S ribosomal protein L35	rpmI	3.36	4.14	-0.78	2.49
Transition state regulatory protein AbrB	abrB	4.44	5.23	-0.79	1.68
Uncharacterized isomerase YfhB	yfhB	2.41	3.32	-0.91	2.77
Ribonuclease PH	rph	1.19	2.11	-0.91	1.83
Sensor histidine kinase ResE	resE	1.39	2.32	-0.93	2.06
Probable ATP-dependent RNA helicase YfmL	yfmL	1.86	2.80	-0.94	2.16
Probable iron uptake system component EfeM	efeM	2.48	3.44	-0.96	1.90
Uncharacterized phosphatase PhoE	phoE	1.53	2.50	-0.97	1.48
Putative L,D-transpeptidase YciB	yciB	2.16	3.16	-1.00	1.48

sp|P17620|RIBBA_BACSU Riboflavin biosynthesis protein RibBA	ribBA	3.71	4.76	-1.05	1.90
MIP18 family protein YitW	yitW	0.33	1.39	-1.06	1.28
Transcriptional regulatory protein ComA	comA	0.33	1.39	-1.06	1.28
Uncharacterized protein YoeB	yoeB	1.72	2.80	-1.07	2.38
Pur operon repressor	purR	2.19	3.27	-1.08	2.23
Elongation factor P	efp	3.84	4.91	-1.08	2.83
Keratin, type II cytoskeletal 2 epidermal OS=Homo sapiens GN=KRT2 PE=1 SV=2	KRT2	2.33	3.41	-1.08	1.31
UPF0296 protein YlzA	ylzA	1.00	2.11	-1.11	3.30
Probable NAD-dependent malic enzyme 4	ytsJ	4.62	5.78	-1.16	3.81
Holliday junction ATP-dependent DNA helicase RuvA	ruvA	1.39	2.57	-1.18	2.10
Fructosamine kinase FrlD	frlD	0.67	1.86	-1.19	1.53
Septum formation protein Maf	maf	0.33	1.53	-1.19	1.27
Uncharacterized transcriptional regulatory protein YvcP	yvcP	0.00	1.19	-1.19	2.44
Uncharacterized hydrolase YxeP	yxeP	0.00	1.19	-1.19	2.44
PtsGHI operon antiterminator	glcT	0.00	1.19	-1.19	2.44
Protein hit	hit	1.39	2.60	-1.21	1.54
Cell division protein SepF	sepF	1.19	2.41	-1.21	2.32
Isoprenyl transferase	uppS	1.33	2.58	-1.25	1.70
Iron(3+)-hydroxamate-binding protein YxeB	yxeB	1.53	2.82	-1.29	1.46
7-cyano-7-deazaguanine synthase	queC	0.67	1.97	-1.30	1.52
Uncharacterized protein YloU	yloU	0.67	1.97	-1.30	1.52
GTP pyrophosphokinase YjbM	yjbM	0.86	2.19	-1.33	1.25
Uncharacterized protein YhbE	yhbE	0.33	1.72	-1.39	1.74
50S ribosomal protein L21	rplU	3.22	4.62	-1.40	2.70
Acetolactate synthase small subunit	ilvH	2.99	4.46	-1.46	3.52
2-hydroxymuconate tautomerase	ywhB	1.19	2.66	-1.46	2.66
General stress protein 16U	yceD	4.09	5.58	-1.49	1.92
Endoribonuclease YbeY	ybeY	1.19	2.71	-1.51	2.24
Uncharacterized protein YqeZ	yqeZ	0.33	1.86	-1.53	1.88
Stage V sporulation protein S	spoVS	0.33	1.86	-1.53	1.88
7-carboxy-7-deazaguanine synthase	queE	0.67	2.21	-1.55	1.94
Cluster of Keratin, type I cytoskeletal 10 OS=Homo sapiens GN=KRT10 PE=1 SV=6 (K1C10_CON-HUMAN)	KRT10	2.00	3.63	-1.63	1.23
Alkaline phosphatase synthesis transcriptional regulatory protein PhoP	phoP	0.00	1.67	-1.67	2.13
Putative aminopeptidase YsdC	ysdC	1.58	3.27	-1.69	5.31
DNA-entry nuclease inhibitor	nin	0.00	1.83	-1.83	2.76
6,7-dimethyl-8-ribityllumazine synthase	ribH	3.84	5.68	-1.85	3.89
Uncharacterized protein YxiF	yxiF	0.00	1.86	-1.86	3.75
2,3-dihydroxybenzoate-AMP ligase	dhbE	0.67	2.57	-1.90	2.21
Uncharacterized protein YbcI	ybcI	1.19	3.11	-1.91	1.32
Putative carboxypeptidase YodJ	yodJ	0.53	2.46	-1.94	1.54
Isochorismatase	dhbB	0.33	2.30	-1.97	2.21
LOG family protein YvdD	yvdD	0.67	2.85	-2.19	2.37
30S ribosomal protein S15	rpsO	3.94	6.15	-2.22	4.58
Dimodular nonribosomal peptide synthase	dhbF	0.86	3.27	-2.41	2.18
Riboflavin synthase	ribE	0.00	2.71	-2.71	3.75

16 - Table 8.10 – Differentially expression proteins between split cluster and wild type 6AA

A list of the all the significantly differentially expression proteins found in our proteomics experiment comparing the split cluster mutant and the wild type in MOPS minimal media supplemented with amino acids. The table includes the gene name and symbol as well as a the log2 counts of the protein peptides, the difference in these values and the negative log2 p-value.

8.3 List of Figures

8.4 List of Tables

8.5 List of Abbreviations

HGT	Horizontal Gene Transfer
TCE	Translation Cell Envelope Cluster
Amp^R	Ampicillin resistance
Cm^R	Chloramphenicol resistance
Erm^R	Erythromycin Resistance
MoClo	Modular Cloning
OD	Optical density
Ori	Origin of replication
TSS	Transcription start site
PCR	Polymerase Chain Reaction
qPCR	quantiative PCR
RFU	Relative fluorescent units
Spc^R	Spectinomycin resistance