Targeting the bacterial cell envelope by molecular coevolution and high throughput phenotyping

Dissertation
zur Erlangung des Doktorgrades
der Fakultät für Biologie
der Ludwig-Maximilians-Universität München

Vorgelegt von
Georgios Kritikos
aus Preveza, Griechenland

München, 2016

Diese Dissertation wurde angefertigt
unter der Leitung von Prof. Dr. Thorsten Mascher
im Bereich von der Fakultät für Biologie
an der Ludwig - Maximilians - Universität München

Erstgutachter: Prof. Dr. Thorsten Mascher
Zweitgutachter: Prof. Dr. Kai Papenfort

Tag der Abgabe: 01.08.2016
Tag der mündlichen Prüfung: 21.11.2016

ERKLÄRUNG

Ich versichere hiermit an Eides statt, dass meine Dissertation selbständig und ohne unerlaubte Hilfsmittel angefertigt worden ist.
Die vorliegende Dissertation wurde weder ganz, noch teilweise bei einer anderen Prüfungskommission vorgelegt.
Ich habe noch zu keinem früheren Zeitpunkt versucht, eine Dissertation einzureichen oder an einer Doktorprüfung teilzunehmen.

München, den 22.11.2016

Georgios Kritikos

"Elegance is not a dispensable luxury
but a quality that decides
between success and failure"
Edsger W. Dijkstra

to my wife

Abstract

Microbial species exhibit a wide repertoire of phenotypic responses to their surroundings, be it stresses posed by their environment, or signals from their bacterial community. Despite advances in computer vision, reporting such phenotypic responses is often done in a qualitative manner. In the course of my work I developed a user-friendly software tool to address the lack of a standardized, quantitative method to measure microbial phenotypes macroscopically. This freely available software, called Iris, can quantify a wide range of microbial phenotypes at the colony level and in a high-throughput fashion. Iris is already used by several research groups, and I present some of its diverse applications and potential for hypothesis generation.

One such application is the quantification of the impact of each gene on the cell envelope permeability in *E. coli*. The Gram-negative bacterial cell envelope forms a barrier against antimicrobial drugs, drastically limiting the list of treatments effective against these organisms. To expand our knowledge on how this multi-layered is built and perturbed, we developed a rapid screening method to detect mutants with envelope defects. By screening a systematic gene deletion mutant collection in *E. coli* across 4 conditions, we identified a number of mutants with defects in envelope assembly. Among those were genes known to be involved in envelope biogenesis, as well as 102 genes of unknown function. In the course of my work I built upon and improved this screening approach, to acquire quantitative membrane permeability measurements that can be used for high-throughput chemical genomics approaches.

Gram-negative bacterial envelope is both a permeability barrier, and a structural barrier. The structural component mainly consists of the rigid peptidoglycan (PG) sacculus, which gives the cells the ability to withstand both turgor pressure and environmental insults. Although biosynthesis of PG is central to bacteria and a target of β-lactam antibiotics, its regulation remains largely elusive. Recently, a number of regulators of PG biosynthesis have been identified, and shown to have coevolved with domains in PG synthases. With the aim of uncovering potential regulatory connections, I developed a computational approach to explore the coevolution of domains in proteins involved in cell wall biosynthesis and remodeling with other proteins in the cell. The method correctly identified existing regulatory interactions, and is readily applied to species across the bacterial kingdom.

Acknowledgements

First and foremost, I would like to acknowledge Nassos Typas who trusted me with learning several analyses techniques, while getting to grips with several facets of microbiology. I am truly grateful for the countless hours it took to mentor me and support me through my PhD, an amazing life-changing journey.

This thesis is dedicated to my wife; without her love and support I wouldn't be writing these very lines. I would also like to thank my parents for their support and loving encouragement during my education path.

My thanks go to *you*, the reader; if I had more time this thesis would be shorter.

Moreover, I wish to thank my two fellow PhD students in Nassos' lab, Lucía and Birgit, for all that we've been through together. These acknowledgements would be incomplete without heartfully thanking all the members in the Typas lab, especially Anja Telzerow and Alexandra Koumoutsi for all their support.

I'm grateful to our collaborators, Carol Gross and KC Huang, for their valuable input and insightful comments. Words fall short when I try to express how grateful I am to Manuel Banzhaf for his constant support, cooking, and mentoring throughout my PhD.

Last but not least, I would like to thank my fellow PhD students across EMBL, who became my family, and my home away from home.

Table of contents

List of abbreviations

ATP	Adenosine triphosphate
BH	Benjamini-Hochberg
BLAST	Basic local alignment search tool
COG	Cluster of orthologous genes
CPR	chlorophenyl red
CPRG	chlorophenyl red β-D-galactopyranoside
DNA	Deoxyribonucleic acid
DUF	Domain of unknown function
GO	Gene ontology
GT	Glycosyltransferase
HMM	Hidden Markov model
IM	Inner membrane
IPTG	Isopropyl β-D-1-thiogalactopyranoside
LPS	Lipopolysaccharide
MI	Mutual information
MSA	Multiple sequence alignment
NCBI	National center for biotechnology information
NLP	Natural language processing
ODD	Outer-membrane PBP1A docking domain
OM	Outer membrane
OMP	Outer membrane porin
PBP	Penicillin binding protein
PG	Peptidoglycan
REST	Representational state transfer
rRNA	Ribosomal ribonucleic acid
SPA	Sequential peptide affinity
STRING	Search tool for the retrieval of interacting genes
Tn	Transposon
TOL	Tree of life
TP	Transpeptidase
TPR	Tetratricopeptide repeat
Und-P	Undecaprenyl phosphate
Und-PP	Undecaprenyl pyrophosphate

List of figures

1 Introduction

Microorganisms were first described in the late 17th century, coinciding with the advent of microscopes. Antonie van Leeuwenhoek first observed life at the microscopic level, while later generations of scientists, like Luis Pasteur, Joseph Lister, and Robert Koch identified the first pathogenic microorganisms. From the discovery of the antibiotic penicillin by Alexander Fleming in 1928 to this day, one of the principal tasks of microbiology has been to understand and combat pathogenic microorganisms. The recent decline in the development of new antibiotics, combined with the increase in multi-drug-resistant bacteria in clinics has set the antibiotic field as a re-emerging priority.

Bacterial cell envelopes hold special interest because of their dual property as both structural, and permeability barriers. Their main structural component, the peptidoglycan cell wall, is exclusively present among bacteria, and thus already the target of many antibiotics. On the other hand, the permeability barrier restricts cell entry for many such compounds.

Bacterial cell envelopes are mainly classified into two major groups: (i) Gram-positive and (ii) Gram-negative cell envelopes. The Gram-positive cell envelope contains two layers: the cytoplasmic membrane (CM), and a cell wall zone containing multi-layered peptidoglycan. Instead, the three-layered Gram-negative cell envelope consists of the CM, the peptidoglycan layer and the outer membrane (OM) (Vollmer and Seligman, 2010). The space defined by the inner and outer membranes in Gram-negative bacteria is termed the periplasm.

1.1 Bacterial envelope

1.1.1 Peptidoglycan cell wall

Peptidoglycan (PG) is the major component of the bacterial cell wall, and is formed by glycan strands that are cross-linked by short peptides. This mesh-like structure gives bacterial cells the ability to withstand their own turgor pressure. Failure to correctly synthesize or maintain peptidoglycan leads to quick bacterial cell death. Moreover, PG is ubiquitously and exclusively present in the bacterial kingdom, thus presenting an excellent target for antibiotics.

Chemically, PG is a polymer of alternating N-acetylglucosamine (GlcNAc) and N-acetylmuramic acid (MurNAc) units. The MurNAc sugars are linked to short peptides of alternating L- and D-amino acids (Schleifer and Kandler, 1972). While the exact composition of the peptide chain can vary significantly across evolution, the peptide sequence in many enterobacterial species is depicted in Figure 1 (Turner et al., 2014). These short peptides of adjacent glycan strands may be connected by the formation of a crosslink most often between in the D-Ala at position 4 of one peptide to the m-Dap at position 3 of another peptide.

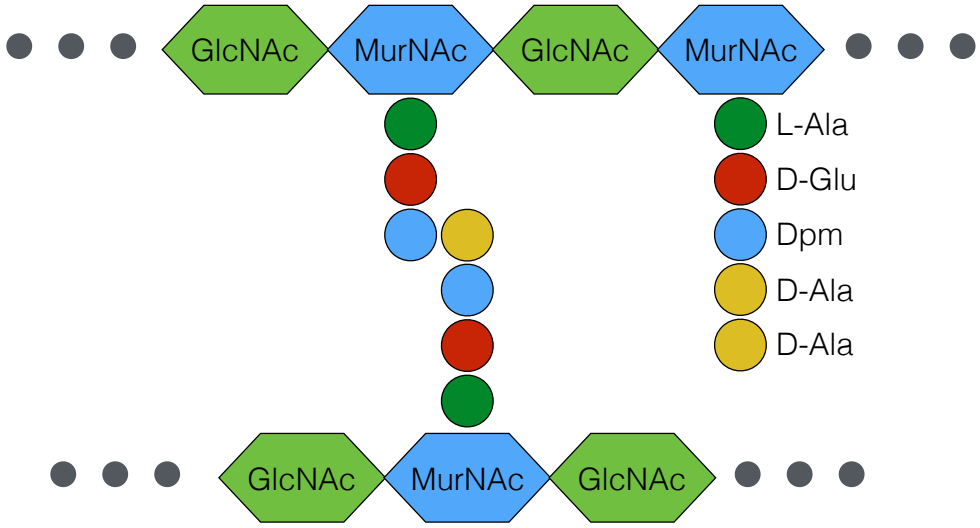

Figure 1: Peptidoglycan cell wall components in many Gram-negative and Gram-positive species. GlcNAc, N-acetylglucosamine; MurNAc, N-acetylmuramic acid; L-Ala, L-Alanine; D-Glu, D-Glutamic acid; Dpm, meso-Diaminopimelic acid; D-Ala, D-Alanine. Adapted from (Turner et al., 2014)

Synthesis

The key enzymes involved in PG biosynthesis are named Penicillin-Binding Proteins (PBPs), discovered by their affinity to the antibiotic penicillin. Bifunctional PBPs, such as the *E. coli* PBP1A, perform elongation of glycan strands, and crosslinking of the strands to form the rigid PG mesh. Glycan strands are formed using the precursor molecule undecaprenyl-pyrophosphoryl-MurNAc-pentapeptide-GlcNAc (lipid-II), which is polymerized by the glycosyltransferase reaction (GT). Peptide cross-links are subsequently formed between adjacent glycan strands by the transpeptidase (TP) reactions (Izaki et al., 1966).

PBP proteins are classified according to the enzymatic functions they perform. Class A PBPs are bifunctional proteins that carry both GT and TP catalytic domains (Goffin and Ghuysen, 1998). On the other hand, class B PBPs also called monofunctional PBPs, and are only able to perform the TP reaction.

Hydrolysis

On the one hand GT and TP functions are key vital steps in PG biosynthesis. On the other hand, controlled strand hydrolysis of the PG mesh is equally as important for PG growth, since degradation is required for inserting nascent material, as well as allowing macromolecular structures, such as efflux pumps, motility, and secretion machineries to cross the PG layer (Scheurwater and Burrows, 2011; Zahrl et al., 2005).

Overall, PG hydrolysis involves several steps performed by different enzyme classes. N-acetylmuramidases, N-acetylglucosaminidases, and Lytic transglycosylases cleave the bonds between GlcNAc and MurNAc in the glycan backbone. N-acetylmuramyl-L-alanine amidases cleave the bond between MurNAc and the first alanine in the peptide chain, removing the peptide chain from the sugar backbone. Carboxypeptidase and endopeptidase enzymes further digest the peptide chain (Typas et al., 2012). Finally, the broken-down PG components are often recovered and reused by a sophisticated recycling mechanism (Park and Uehara, 2008).

Regulation

While PG hydrolysis is important for cell homeostasis and function, uncontrolled hydrolysis will lead to rupture of the PG sacculus and cell lysis. This happens upon inhibition of PG biosynthesis, since PG hydrolysis continues. Therefore, hydrolase activity is thought to be controlled by incorporation into large protein complexes of PG synthases and hydrolases. In this way, hydrolases would be localized in sites of PG synthesis, which would also guide removal of old PG material, a step necessary for incorporation of new material (Höltje, 1998). However, evidence of such large multi-enzyme protein complexes remains elusive to this day.

On the other hand, macromolecular complexes were recently shown to play an important role in the regulation of PG synthases (Typas et al., 2012). For both major PG synthases of *E. coli*, PBP1A and PBP1B, two outer membrane lipoproteins have been shown to bind to their cognate PG synthase and to stimulate their activity. In order to do so, the lipoproteins LpoA and LpoB need to

traverse the PG sacculus to physically bind to PBP1A and PBP1B respectively. This lipoprotein-mediated activation of PG synthesis is considered to provide a way to control the PG layer thickness, since a thick PG layer would prevent access of the Lpo activators to their cognate PG synthase (Typas et al., 2012). Lpo proteins need to pass through pores in the PG mesh to interact with their cognate PBP. In higher turgor pressure cells, such as fast-growing cells, the increased pore size would allow for better access of the Lpo to the cognate PBP. This would in turn elegantly couple the cell growth rate to the PG biosynthesis rate.

1.1.2 Outer membrane in Gram-negative bacteria

The outer membrane (OM) of Gram-negative bacteria presents a formidable barrier that antibiotic compounds need to overcome in order to reach their molecular targets (Delcour, 2009; Silhavy et al., 2010). Like other biological membranes, the OM is a lipid bilayer. However, the OM is asymmetric, with the outer leaflet consisting predominantly of lipopolysaccharide (LPS), while the inner leaflet consists mainly of phospholipids.

LPS is essential to the barrier function of the outer membrane limiting the diffusion of hydrophobic molecules such as detergents, and bile salts (Wang and Quinn, 2010). The OM is crossed by β-sheet proteins assembled into cylinders and are referred to as Outer Membrane Porins (OMPs). The majority of OMPs facilitate the diffusion of small molecules across the OM, with a limit at around 700 Daltons (Nikaido, 2003). Combined with the fact that LPS molecules provide a very effective barrier for hydrophobic molecules, this makes the OM a very selective permeability barrier (Silhavy et al., 2010). This barrier function also shields Gram-negative bacteria against antibiotic compounds, drastically limiting the repertoire of antibiotics effective against these organisms (Delcour, 2009).

Importantly, modifications to the LPS and OMPs have often been observed as survival strategies of pathogens against antibacterial molecules. For example, *Salmonella enterica* serovar Typhimurium is known to increase the density of its LPS to decrease the influx of host cationic molecules in the context of invasion (Delcour, 2009). On the other hand, several species such as *Klebsiella pneumonia*, *Pseudomonas aeruginosa, and Neisseria gonorrhoeae* have been reported to acquire antibiotic resistance through modifications to their OMPs.

In *E. coli* the OM is tethered to the rigid PG layer through the highly abundant Lpp protein (Braun and Wolff, 1970). Moreover, the OM and PG layer were shown to coordinate their constriction via the formation of a macromolecular complex linking the OM-constricting Tol system to PBP1B, the bifunctional PBP specialized in cell division (Gray et al., 2015).

1.2 Quantitative phenotypic profiling: discovery of gene function and pathway organization

The number of sequenced bacterial genomes has exploded in the last decade. Currently, almost 5,000 prokaryotic genomes are completely sequenced and assembled (NCBI, Jan 2016). At the same time, gene function annotation is lagging behind. Even in the well-studied bacterium *E. coli*, 1600 genes (37% of its genome) remains of unknown function (Hu et al., 2009), (Biocyc, Jan 2016).

In the past decade, the elucidation of gene functional relationships, as well as complex and pathway identification was greatly aided through the advent of high-throughput reverse genetics approaches. Such approaches include gene-drug and gene-gene (chemical genomic and genetic interaction) screens that were first pioneered in S. cerevisiae (Parsons et al., 2006; Roemer et al., 2012; Tong et al., 2004), and later expanded on multiple microbial species, including bacteria (Brochado and Typas, 2013; Deutschbauer et al., 2011; Nichols et al., 2011; Pasquina et al., 2016; Phillips et al., 2011; Typas et al., 2008).

Such approaches rely on the measurement of a quantitative phenotype, either at the microscopic cell level, or at the macroscopic population level. Cell-level approaches are typically based high throughput microscopy, whereby multiple features for every microscopy image is extracted and analyzed (Ohya et al., 2005; Styles et al., 2016). On the other hand, macroscopic population-level phenotypic studies often measure growth fitness on solid surface (Hillenmeyer et al., 2008; Parsons et al., 2006; Tong et al., 2004)

However, measuring only growth fitness was shown to limit the phenotypic characterization potential of such screens. For example, in a recent such screen in *E. coli*, more than 3700 gene deletion strains (Baba et al., 2006) were queried against more than 320 conditions (Nichols et al., 2011). Despite the wide range of stresses, about half of gene deletion strains showed no statistically significant

growth phenotype. These genes were enriched in processes such as pili biosynthesis, chemotaxis and biofilm formation, all of which are unrelated to growth fitness.

At the same time, there are several established assays for phenotypes in processes uncoupled from growth fitness. Such assays include biofilm formation, colony morphology, and sporulation. Performing high-throughput phenotypic screens with such assays would thus provide a novel systematic way of exploring such processes, and thus new insight into the biology of assayed species.

2 Iris: Expanding the palette of microbial phenotypic readouts

2.1 Summary

In the course of my thesis I developed a versatile image analysis software to quantify a number of different microbial colony phenotypes. The software is called Iris and has already numerous users taking advantage of its many applications across diverse microbial species. In the present chapter, I briefly supply the motivation for developing such software, as well as a quick overview of its applications. In most projects where Iris was used, I was also heavily involved in data analysis. The results of several such projects have been published, submitted for publication, or are soon to be submitted. I list all such manuscripts below. Unless otherwise stated, my contribution to these projects amounts to the development of phenotype quantification and subsequent data analysis. At the same time, the experimental part of these projects was performed by colleagues in the Typas lab or in collaborating labs.

1. Paradis-Bleau C, **Kritikos G**, Orlova K, Typas A, Bernhardt TG (2014) A Genome-Wide Screen for Bacterial Envelope Biogenesis Mutants Identifies a Novel Factor Involved in Cell Wall Precursor Metabolism. PLoS Genet 10(1): e1004056. doi: 10.1371/journal.pgen.1004056

2. Shiver A, Osadnik H, **Kritikos G**, Li B, Krogan N, Typas A, Gross C (2016) A Chemical-Genomic Screen of Neglected Antibiotics Reveals Illicit Transport of Kasugamycin and Blasticidin S. PLoS Genet 12(6): e1006124. doi: 10.1371/journal.pgen.1006124

3. Koo BM, **Kritikos G**, Farelli J, Todor H, Tong K, Kimsey H, Wapinski I, Galardini M, Cabal A, Peters J, Hachman A, FitzGerald M, Hung D,

Rudner D, Allen K, Typas A, Gross C (submitted to Cell Systems, currently in review) Building Systems Resources for the Model Gram-positive bacterium Bacillus subtilis.

4. **Kritikos G**, Banzhaf M, Herrera L, Koumoutsi A, Typas A (in preparation) Iris: expanding the palette of microbial phenotypic readouts.

2.2 Background and significance

Defining a microbial colony phenotype has traditionally been a qualitative, descriptive process. Although easy to communicate, such qualitative traits are hard to compare, record, and obtain systematically. Such qualitative descriptions are also often subject to human error, as well as biases related to expected phenotypic outcome.

The ability to automatically quantify colony phenotypes alleviates such issues, in addition to bringing about several other advantages. For example, one can use these quantitative values for assessing phenotype reproducibility, as well as the statistical significance of a phenotypic deviation from a population. Moreover, a standardized quantification tool could readily be used to compare results across studies.

Perhaps most important of all, quantitative data on microbial phenotypes can readily be used for large-scale phenotyping. High throughput gene-drug and gene-gene (chemical genomic and genetic interaction) screens were first pioneered in *S. cerevisiae* (Hillenmeyer et al., 2008; Roemer et al., 2012; Tong et al., 2004), and later expanded on multiple microbial species, including bacteria (Brochado and Typas, 2013; Deutschbauer et al., 2011; Nichols et al., 2011; Pasquina et al., 2016; Phillips et al., 2011; Typas et al., 2008).

Such approaches rely on automatic quantification of a colony phenotype to assess the interaction. Genetic and chemical-genetic interaction screens have led the way towards the elucidation of gene functional relationships in the past decade, as well as identifying complexes and pathways.

In a recent such screen in *E. coli*, more than 3700 gene deletion strains (Baba et al., 2006) were queried against more than 320 conditions (Nichols et al., 2011). Despite the wide range of stresses, a large number of gene deletion strains showed no statistically significant growth phenotype. These genes were enriched in

processes such as pili biosynthesis, chemotaxis and biofilm formation, all of which are unrelated to growth fitness.

At the same time, a large number of genes of unknown function also failed to give a significant growth phenotype. Since measuring only growth phenotypes was not sufficient to have a complete picture of the biology of a species, it became clear that future chemical genomics approaches need to measure more phenotypes.

In the lab we undertook the task of devising assays that would capture new phenotypes to provide information orthogonal to that of growth fitness. Such assays included biofilm formation and sporulation. Currently, a range of available software can measure microbial colony size (Memarian et al., 2007; Wagih and Parts, 2014; Wetherow et al., 2010), but there was no available software that can quantify microbial colony phenotypes other than colony size. As an example, recent studies using colony morphology as readout rely on manual semi-quantitative phenotypic characterization (Cabeen et al., 2016; Ryan et al., 2012). To pair these high-throughput assays with a quantitative readout, I developed an automated image analysis software for automated phenotype quantification.

Colleagues such as Lucia Herrera and Manuel Banzhaf undertook the task of optimizing assays such as the congo-red staining biofilm formation assay for their use in chemical genomics approaches. In parallel to developing a fully automated image analysis software to allow for complex phenotype extraction, I was involved in the optimization of the many of the new quantitative high-throughput screens.

2.3 Goals

My goal was to develop an open-source software tool that can accurately quantify microbial colony phenotypes that go beyond colony size. At the same time, the software needs to be fully automated, and easy to use by non-experts.

Most importantly, the software needs to be designed in a way that is easy to expand its potential applications. To achieve this, the Iris source code was designed in a way that allows for easy extension of its functionality. This can be done either by mixing and matching existing readout functions or by developing new ones. The last goal was proven to be most useful, since the applications of Iris have expanded with the number of users. Below I list some of the applications, and also the diverse biological questions that led to them.

2.4 Results and applications

2.4.1 Image processing pipeline

High throughput phenotypic assay quantification starts with a typically high-resolution picture of a colony array on a rectangular agar plate (see Figure 2). The first processing step for the software is to automatically rotate the picture so that the colony array is perfectly horizontal. In the next step Iris detects and crops the plate boundaries. A cropped picture containing only the colony array is then segmented into picture tiles, each holding only one colony. Each tile is then separately processed by one or several tile processor modules, each specialized quantifying a specific phenotype (e.g. sporulation). This design allows for independent colony quantification, but also easy incorporation of new readouts.

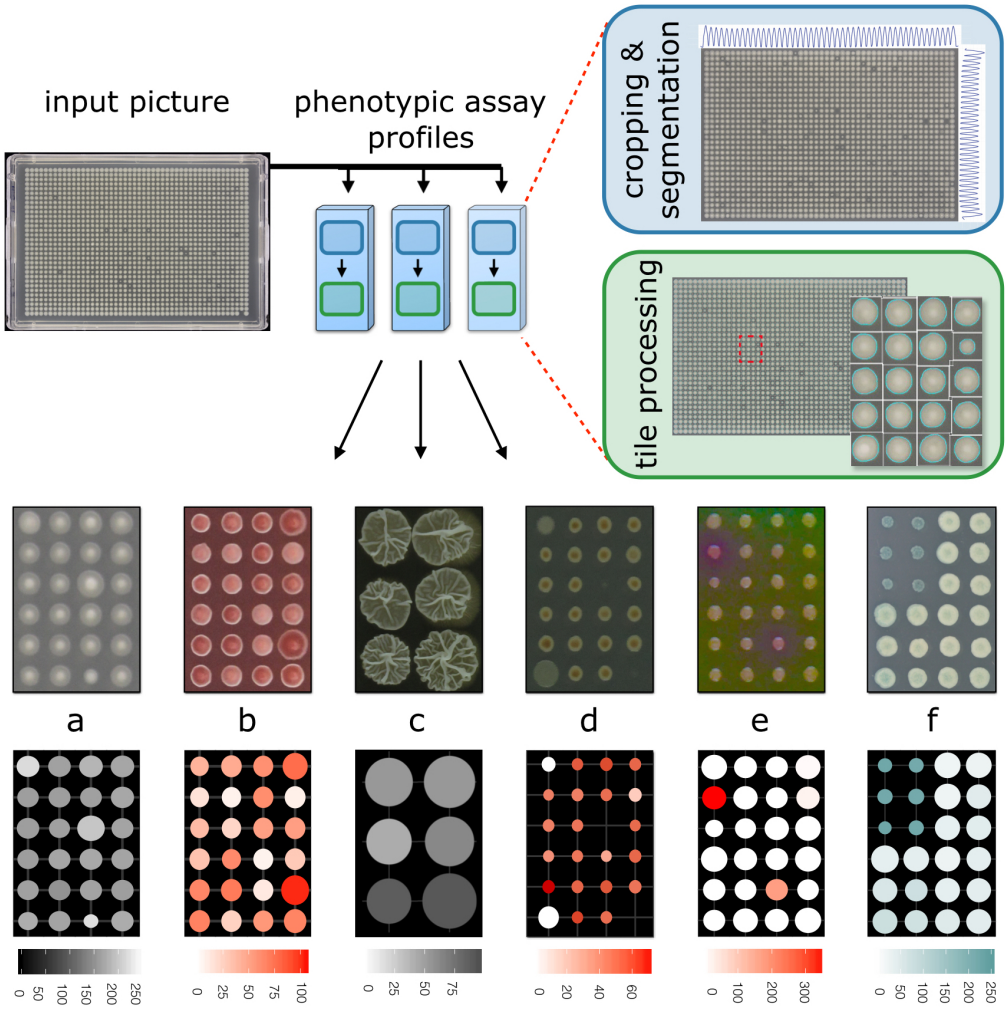

Figure 2: overview of Iris software design and image processing pipeline. Bottom: examples of different phenotype quantification profiles: a. colony opacity (*E. coli*), b. biofilm formation (*E. coli*), c. colony morphology (*C. albicans*), d. sporulation (*B. subtilis*), e. membrane permeability (*E. coli*), f. β-galactosidase activity (*E. coli*)

Iris is designed in a modular fashion to allow for easy extensibility to new assays and readouts. Each processing task in Iris is performed by a separate module, while different modules performing the same task are easily interchangeable. This allows for expert users to write custom made modules to fit the needs of assays not covered already by the distributed Iris version.

2.4.2 Colony opacity

Across high throughput chemical genomics screens, colony size (area) is used as a proxy for mutant growth fitness. However, microbial colonies are known to grow in area, but also in height and density. Despite commonly used, simple size measurement does not reflect colony height nor density, thus lending itself to error approximating growth fitness.

Iris on the other hand also measures colony opacity, which takes into account colony area, height and density. Colony opacity is measured by summing the per-pixel over-background brightness values for all the pixels in the colony bounds. The over-background brightness for every pixel is in turn calculated by subtracting the pixel brightness to the average brightness of background pixels.

By measuring colony size and opacity at the same time, I was able to detect mutants that form either very translucent, or very dense colonies. For such mutants, colony size would either over- or under-estimate their growth fitness. On the other side of the spectrum, very dense colonies can also be due to the secretion of extracellular material. For example, colanic acid is known to yield this bright mucoid material that covers the cells and thus permeates the colony. Such outliers in density do not reflect actual growth fitness but a different biological phenomenon altogether.

2.4.2.1 Comparing colony opacity to colony size

Using Iris I reanalyzed the original pictures acquired for a large *E. coli* chemical genomics dataset (Nichols et al., 2011). I thus acquired both colony size and colony opacity information for more than 3700 *E. coli* mutant strains across more than 320 conditions.

Using the newly-acquired opacity dataset, I then followed the same statistics analysis as in the original study. For most mutants across conditions colony opacity readout closely follows colony size. Interestingly, treatment with inhibitors of fatty acid biosynthesis, was found drastically increased colony density, implying that all these stresses target *E. coli* more when it reaches stationary phase or affect the colony architecture. In contrast, conditions such as A22 that induce colanic acid production (a mucoid secreted polysachharide) (Cho et al., 2014), result in colonies with higher density (Figure 3b). Many mutants also exhibited condition-

specific phenotypes, such as YciB, a poorly characterized protein with a proposed role on cellular morphogenesis (Li et al., 2015), and FadD, the long-chain fatty acid acyl-CoA synthetase, which plays a pivotal role on transport of exogenous fatty acids in the cell (see Figure 3c).

I then built upon this new fitness readout to acquire fitness scores (S-scores) for every mutant across each condition as previously described (Nichols et al., 2011). By correlating the vector of S-scores for each mutant across conditions, we acquired a gene association network which we compared with the equivalent published network calculated using colony size S-scores (Nichols et al., 2011). Comparison of the two datasets, revealed a number of mutants for which we gained association power to other mutants and pathways (see Figure 3d). As an example, molybdopterin synthesis pathway genes feature higher correlations to genes within the pathway, as well as with genes in related processes (see Figure 3e).

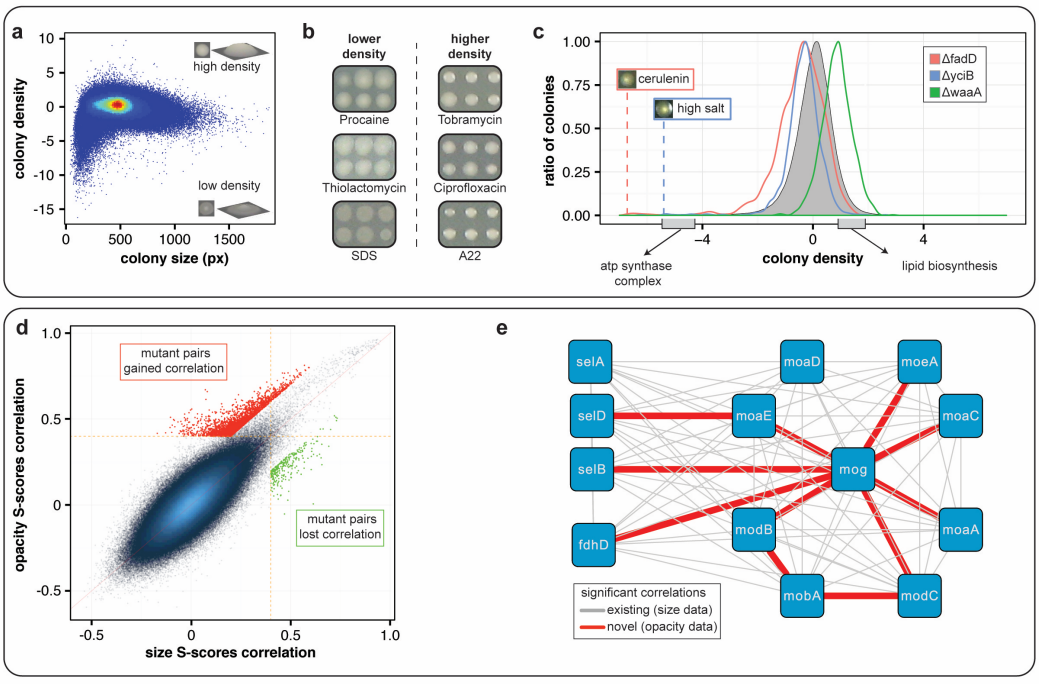

Figure 3: Reanalyzing the data of a large-scale chemical genomics screen using colony opacity as a growth fitness proxy. Top panel: colony density; a. acquiring colony density allows for better resolution at severe growth defects (small size), b. chemical conditions feature distinct reactions in colony density, such as A22; c. density phenotypes hold information not available by size alone, such as conditional phenotypes in *ΔfadD* and *ΔyciB* strains. Bottom panel: d. re-analysis of fitness data (Nichols et al., 2011) using colony opacity (product of density and size) yields more significant correlations among phenotypic signatures of related genes; e. opacity-derived phenotypic signature correlations capture more interactions (shown in red) among genes coding for molybdenum cofactor biosynthesis and related enzymes.

2.4.2.2 A Chemical-Genomic Screen of Neglected Antibiotics Reveals Illicit Transport of Kasugamycin and Blasticidin S

Colleagues in the Gross lab (UCSF, CA, USA) recently performed a small chemical genomics screen in which the *E. coli* single gene deletion library (Baba et al., 2006) was screened against 30 antibiotics. The antibiotic compounds selected among had either poorly characterized mode of action or resistance mechanisms, or were disused compounds. The goal of the study was to better characterize these drugs, and elucidate their resistance determinants.

Shiver et al. leveraged the power of existing chemical genomics datasets in *E. coli* (Nichols et al., 2011), by merging their data to the existing dataset. Reanalysis of both the existing and the new datasets with Iris using the opacity readout greatly helped dataset integration (Shiver et al., 2016).

This led to the investigation of kasugamycin and blasticidin S mode of action. Both compounds are translation inhibitors and elicited a similar reaction across *E. coli* mutants. Shiver et al. showed that both antibiotics make use of the peptide ABC-transporters Opp and Dpp to promote their entry into bacterial cells (see (Shiver et al., 2016)).

In this study, I provided input related to the reanalysis of the existing chemical genomics dataset. I also aided with the integration of existing and new dataset. Moreover, I provided assistance and software for steps ranging from data acquisition to statistical analysis of the results.

2.4.2.3 Kinetics data

A recent study explored growth kinetics of *E. coli* colonies using document scanners (Takeuchi et al., 2014). A result of that study was that measuring kinetics of the opacity at the center of the colony provides for a more accurate proxy of growth fitness.

The Iris software is accompanied by software that can calculate and visualize mutant growth curves. This open-source companion software is based in the R programming language (Kahm et al., 2010; Team, 2015) and is freely distributed together with the Iris software.

All of the aforementioned Iris features for microbial colonies can also be followed up over time, by means of the provided companion R framework. I expect such a feature to be especially useful for bacterial social phenotypes, such as biofilm formation. As an example of such an application, I track 3 fitness-related colony features over time for all mutants in the E. coli KEIO library (Baba et al.,

2006) (see Figure 4a). Beyond the already described colony area, and colony opacity, Iris also reports colony center density, a fitness measure recently demonstrated to be advantageous when tracked over time (Takeuchi et al., 2014).

I found that the accuracy of the fitted Gompertz model depends on the ability to capture early timepoint values. To improve the accuracy of our models, I tested combinations of different media and lighting conditions to find the ones allowing for earlier colony detection. I found that addition of a dye in the media (Congo Red) provides with better contrast, and thereby leads to robust colony detection as early as 1 hour post-inoculation (see Figure 4a). Using this lighting and media combination, I used the Gompertz model to fit the time series of all 3 fitness measures and calculate lag phase, slopes, and estimated maximum growth values (see Figure 4b).

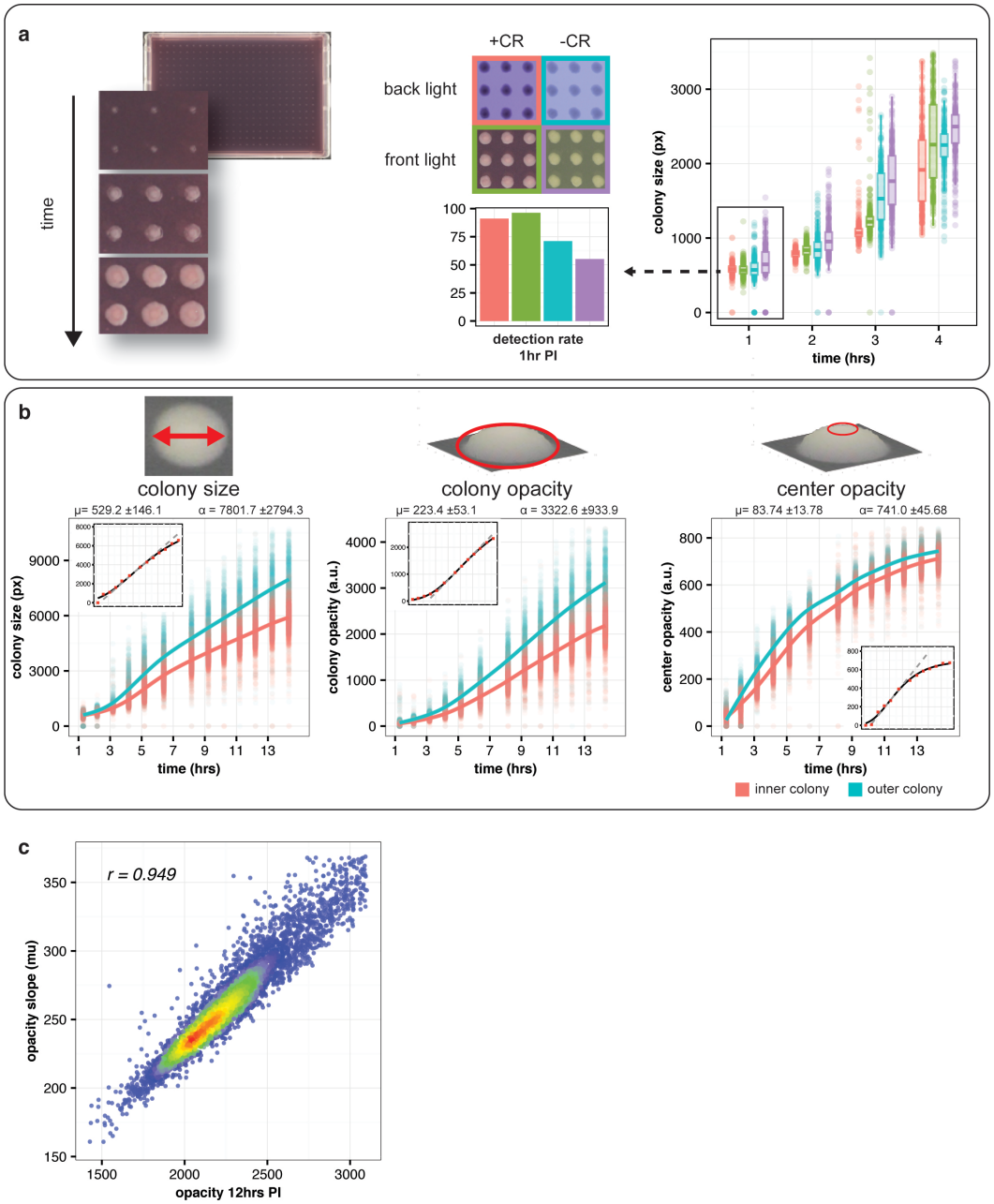

Figure 4: Using Iris for growth kinetic measurements. A: Example of a KEIO *E. coli* deletion strain collection plate arrayed on 1536 format in a CR+ plate photographed using front lighting over different timepoints. B: colony detection rate at 1hr post-inoculation varies according to media and lighting condition used. Front light and congo red (CR) was found to have the best detection rate. C: growth curves using different readouts, demonstrating differences between inner (red) and outer (blue) rows and columns of a screening plate. Numbers depict Gomperz fit parameter average and SD across the whole KEIO library: α corresponds to predicted plateau, while λ corresponds to maximum slope per mutant growth curve. Inlets show example colony growth curves with average slope, as well as the Gompertz fit and maximum slope. D: opacity values for a single timepoint (12hrs post-inoculation) compared to opacity slopes (using timepoints 1-14hrs) across all KEIO collection mutants

Moreover, I compared Gompertz slope (mu) values and end point values for all described measures of mutant fitness. We found that a carefully selected single-

timepoint measurement can be a very good proxy of growth curve slope (Pearson r=0.949, see Figure 4c). In conclusion, the provided framework alongside the Iris software greatly simplifies the process of kinetic data analysis and visualization for all readouts the Iris software provides. Using this framework, we observed that endpoint fitness measurements of *E. coli* mutants provide a good proxy of growth rates.

2.4.3 Colorimetric assays

A wealth of microbial assays relies on color observation to capture phenotypes unrelated to growth. Colorimetric assays typically measure dye or chromophore concentration in or next to a microbial colony as a result of a particular phenotypic trait (biofilm, sporulation, cell lysis, etc). Such assays can be used in high-density colony arrays to phenotype whole microbial libraries (Paradis-Bleau et al., 2014).

The Iris software can readily quantify colorimetric assays including membrane permeability (Paradis-Bleau et al., 2014), B. subtilis sporulation (Byoung-Mo et al., submitted for publication), biofilm formation (Herrera et al., Banzhaf et al., in preparation), and reporter activity (e.g. β-galactosidase).

I am describing the membrane permeability assay results and analysis in a separate chapter, since its development and optimization became a large part of my work.

2.4.3.1 A Genome-Wide Screen for Bacterial Envelope Biogenesis Mutants Identifies a Novel Factor Involved in Cell Wall Precursor Metabolism

Gram-negative bacterial cell envelope forms a barrier against antimicrobial drugs, drastically limiting the list of treatments effective against these organisms. Exploring phenotypes related to cell envelope biosynthesis could lead to new mechanistic insights into this vital process, as well as new drug targets.

In this study, Iris was used to detect a chromophore reaction involving Chlorophenol red-β-D-galactopyranoside (CPRG). CPRG is a galactoside analog and is hydrolyzed into the red chromophore CPR by the cytoplasmic enzyme β-galactosidase. However, CPRG does not penetrate *E. coli* cells if their cell envelope

is intact. Chromophore reaction only takes place only in mutants with membrane biosynthesis defects or with higher lysis frequency, thus colonies of such mutants turn red with time.

A genome-wide *E. coli* gene deletion library (Baba et al., 2006) was screened across 4 conditions, and membrane permeability values were automatically quantified using Iris (see **Figure 5**). Among strains positive for the chromophore reaction were mutants of genes known to be involved in cell envelope biogenesis. The screen also captured 70 genes, mutants of which had elevated lysis phenotype but no growth phenotype captured by the previous larger chemical genomics screen (Nichols et al., 2011). As proof-of-principle, one of those genes, elyC, was further characterized to encode for a factor playing a critical role in the metabolism of the essential lipid precursor required for cell wall biosynthesis.

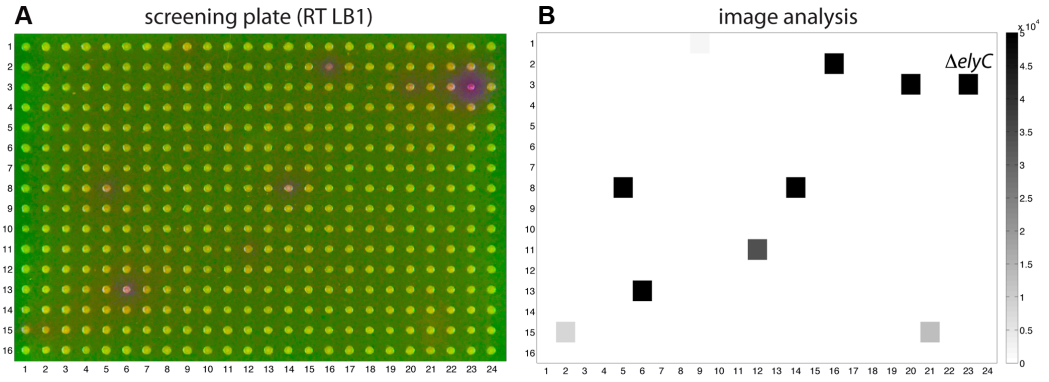

Figure 5: Example of Iris quantification of membrane permeability by detecting the CPRG chromophore reaction. A. Example of a screening plate at room temperature with 1% NaCl. B. Iris-reported CPRG values for each of the colonies in the plate, ΔelyC mutant is located on the top right. Figure adapted from (Paradis-Bleau et al., 2014)

Iris detects the chromophore reaction by detecting the red color hue within the bounds of the bacterial colony. To avoid chromophore diffusion from neighboring colonies, early timepoints were used. Chromophore concentration in early timepoints is low even in positive control colonies. In order to provide for more sensitive chromophore detection at an early timepoint, Iris converts the picture to the cylindrical HSV color space (Smith, 1978). Subsequently, red hue is quantified as the radial distance between the hue of every pixel and the hue of the red chromophore color (0°), and summed across colony pixels.

My contribution to the above study was to expand Iris functionality to provide for a sensitive detection of the chromophore reaction. Moreover, I analyzed Iris data to produce the list of genes positive for chromophore turnover.

I also developed software to perform the functional enrichment analysis for the list of genes positive for chromophore turnover.

2.4.3.2 Building Systems Resources for the Model Gram-positive bacterium *Bacillus subtilis*

Gram-positive bacteria are interesting because of their adaptability to extreme and diverse niches. *B. subtilis* is the model organism in Firmicutes and Gram-positive bacteria in general. Several powerful genetic and cell biology tools have been developed for *B. subtilis*, allowing researchers to study a number of its key core cellular processes, including developmental programs, like biofilm, competence, and sporulation.

The phylum of Firmicutes is especially important, since it is one of the two most abundant phyla in the gut microbiome (Arumugam et al., 2011; Ley et al., 2008; Turnbaugh et al., 2009), while their abundance has been linked to disease (El Feghaly et al., 2015; Miquel et al., 2013). However, with the exception of the minimal genome bacterium *Streptococcus sanguinis* (Xu et al., 2011), there has been no systematic genome-wide approach to elucidate gene function in any Gram-positive organism.

Koo et al. constructed two single-gene deletion libraries of the model Gram-positive bacterium *B. subtilis*, and used them to establish high-throughput screening methodologies for this bacterium. Thereby, *B. subtilis* gene essentiality, auxotrophy, competence, and sporulation were addressed for the first time in a genome-wide fashion. While the acquired data is a valuable resource to the field, the libraries themselves are an even more valuable, versatile resource for the study of gene function.

My contribution to this study was to expand the Iris functionality to acquire fitness values for *B. subtilis*. Since *B. subtilis* often forms quite translucent colonies, I adapted the opacity readout (see 2.4.2) to provide for an accurate proxy of growth fitness.

I also developed a new colorimetric readout for the detection of sporulating cells in *B. subtilis* colonies and lead the data analysis for this screen. Sporulation is the major developmental pathway of *Bacilli* and *Clostridia*, and an important infection strategy in pathogenic species of these genera. In *B. subtilis*, the key sporulation-related transcription factors, regulons, and inter-compartmental communication strategies have been mapped and studied for decades (Higgins

and Dworkin, 2012). In total, nearly 25% of the genome (~1000 genes) have been implicated to take part on the implementation of this key stress developmental program (Eijlander et al., 2014). However, a genome-wide quantitative assessment of the impact of each gene on the sporulation program has never been done. Based on the fact that sporulating cells turn dark brown in minimal medium at a late stage of spore development (Driks, 1999), we (Koo et al.) developed a high-throughput methodology to assess sporulation (see Figure 6a), and used it to determine the relative contributions of known and new sporulation contributors in our two libraries.

The sporulation assay was reasonably reproducible as biological replicates within the same library were highly correlated (Pearson r=0.88, (see Figure 6b) whereas the correlation somewhat dropped (Pearson r=0.68) when comparing the two libraries, mostly due to a small fraction of clone discrepancies. Nevertheless most known sporulation mutants exhibited consistently low sporulation scores (see Figure 6c). We removed poorly-grown mutants from subsequent analysis, because growth affected our ability to accurately quantify pigment development. Non-reproducible mutants between libraries turned out to be due to a rare point mutation on a sporulation key gene that a tiny fraction of the library carried, and mutants were cleaned from the libraries. In total, of the 101 known sporulation mutants present in the filtered data, we recovered 71 at 5% FDR and 79 at 10% FDR. We did not expect to recover 10 of the remaining 22 mutants either because they either had media specific defects or because they were involved in producing dipicolinic acid for spore heat resistance, a step that is independent of pigment development. We recovered genes encoding the quorum-sensing peptides *phrA* and *phrE*, transcriptional regulators and sporulation sigma factors, the sporulation phosphorelay, the *spo0A* phosphorylation stimulating complex, and most spo genes. Notably, *phrA* and *phrE* mutants would be lost in a pooled screen as they would be complemented by the predominantly wildtype cells in the population (Meeske et al., 2016). Gene Set Enrichment Analysis (GSEA (Subramanian et al., 2005)) of sporulation scores indicated that low scores are enriched in cell cycle/division, translation, cell envelope biogenesis, signal transduction and cell motility functional categories ($p<0.05$, see Figure 6d). Of these, the cell motility category was surprising, and we noted that enrichment reflected a general shift to lower sporulation values rather than specific poorly sporulating mutants. Using a stringent 5% FDR as a cutoff ($SS \leq 0.31$), we identified 33 poorly characterized genes as having a dramatic effect on sporulation development. Among those

genes, *ywmB* and *yqzE* were very recently identified as a mother-cell activator of SigE and a forespore activator of SigG, respectively (Meeske et al., 2016). Additionally, 40 genes of known function with a role in sporulation were identified (5% FDR), and 12 of them have been validated (Meeske et al., 2016).

Finally, we used GSEA to assess the contribution of genes in the known sporulation regulons to sporulation. Sporulation is governed by initiation factor Spo0A and a subsequent hierarchical cascade of sporulation sigma factors, SigE, F, K and G (Eijlander et al., 2014). As expected, the sporulation defective phenotype is enriched in these regulons ($p<0.05$) except for SigG (see Figure 6d) where lack of enrichment may reflect its regulation of primarily late-acting proteins. Interestingly, only 52/407 genes, or 10-20% of the genes in each regulon, are significantly sporulation defective (SS <0.31), thereby extending the conclusions reached by a previous study of the SigE regulon (Eichenberger et al., 2003) that the sporulation program encompasses significant redundancy to achieve robustness. Notably, genes expressed in the mother cell are just as likely as those expressed in the forespore to have severe phenotypes, indicating the important role of the mother cell in nourishing and orchestrating forespore development. Mutants of sporulation sigma factor regulons with no or weak phenotypes may have subtle or redundant effects in tested conditions. This study points out the key regulon members that need to be placed in the sporulation-wiring diagram, and also indicates those genes requiring additional dissection, possibly by double mutant analysis (Silvaggi et al., 2004).

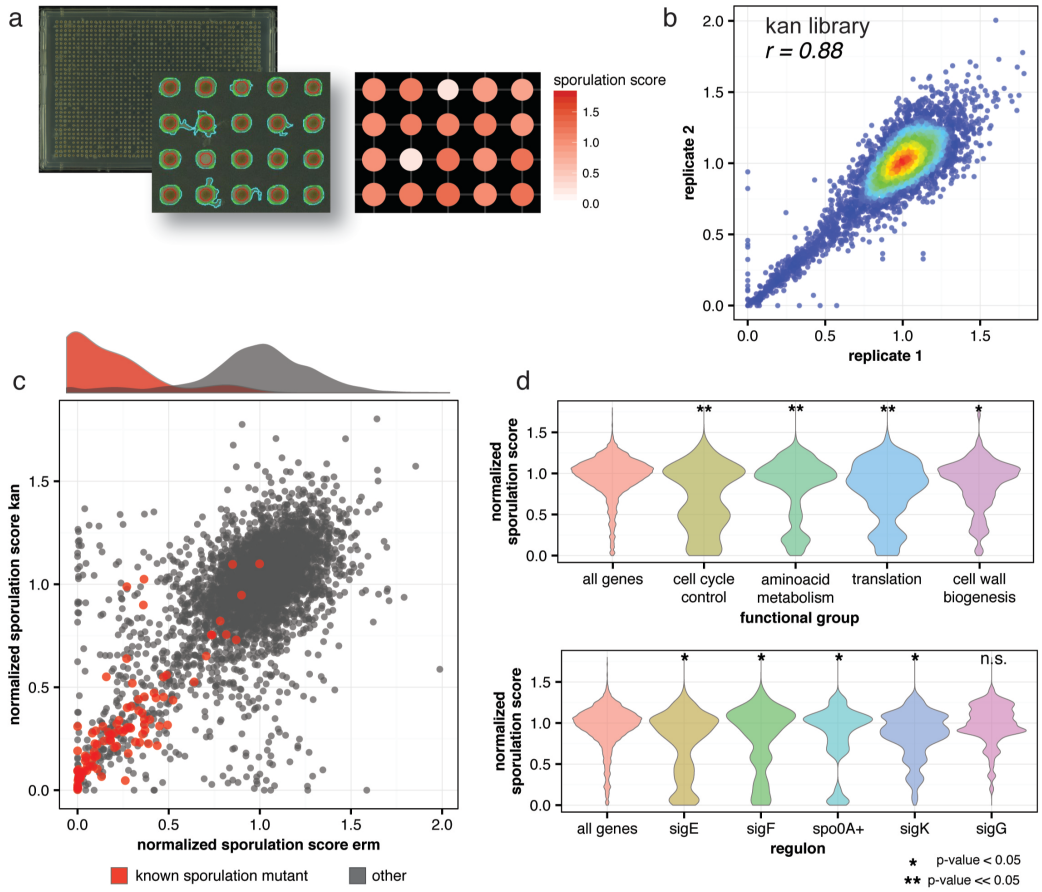

Figure 6: Systematic analysis of sporulation defects in two *B. subtilis* ordered deletion library mutants. (a) Sporulation phenotype and colony size are automatically quantified using Iris after 45 hrs of growth on succinate-glutamate minimal agar plates supplemented with limiting amounts of nutrients. A sporulation plate image is shown at the top. Zoomed part of 1536 colony array image processed by Iris are shown at lower left. The color intensity in the center area of colony (in red circle) is processed for calculating raw sporulation score. Color-coded sporulation scores (SS) of mutants are shown on the right. (b) Reproducibility of SS from two technical replicates of KanR library. (C) Density (top) and scatter (bottom) plot comparing SS of ErmR and KanR mutants. Red color indicates sporulation scores of 111 known sporulation-defective mutants. Using a 5% false discovery rate (FDR), 70% of known sporulation mutants was recovered in this screen. (d, top) Functional groups enriched in sporulation defective mutants (p<0.05). Distribution of SS of genes in each functional category is shown by violin plot. (d, bottom) Distribution of rSS of genes that are positively regulated by mother cell sporulation sigmas (SigE and SigK, colored in red), forespore sgma (SigF and SigG, colored in blue) and Spo0A (colored in yellow). Sigma regulons enriched in sporulation defective genes are indicated by *, p-value <0.05); n.s., not significant

My part was to develop a detection method in Iris to robustly and accurately quantify the degree of sporulation in *B. subtilis* (see 6.1.3.3). Subsequently, I analyzed both sporulation and fitness data for all mutants, and treated the data as in previous approaches to remove systematic, technical artifacts. I analyzed the treated data to calculate reproducibility within and between libraries. I also provided software and analysis related to quality control, such as filtering out

mutant phenotypes that are discordant among the two libraries. Finally, I performed the functional enrichment analysis based on the sporulation phenotype data. Results of this analysis are submitted for publication (Koo et al. in review).

2.4.3.3 Biofilm chemical genomics in *P. aeruginosa* and *E. coli*

Biofilm formation is a key lifestyle decision for bacteria, activated in response to environmental cues, as well as cell-to-cell signaling such as quorum sensing (Hall-Stoodley et al., 2004). Importantly, bacteria growing in a biofilm are shielded against a wide range of antimicrobial compounds. Understanding biofilm development and elucidating its key genetic elements in pathogenic organisms such as *P. aeruginosa* is of paramount clinical importance, since biofilm-growing cells are known to persist in medical equipment as well as in patients (Donlan and Costerton, 2002; Hall-Stoodley et al., 2004).

Banzhaf et al. developed a high-throughput method to assay *P. aeruginosa* mutants for biofilm development and growth fitness. In parallel, Herrera et al. developed the equivalent high-throughput assay for *E. coli* mutants. Both bacterial species were independently assayed for biofilm formation and growth fitness in a wide range of chemical perturbations.

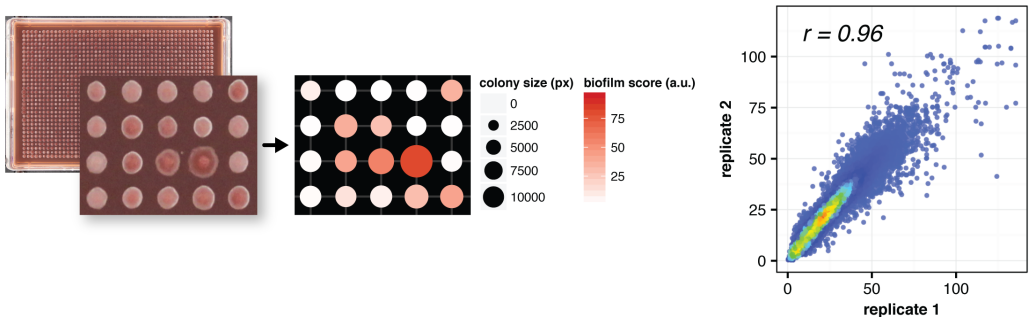

Figure 7: Congo red biofilm formation assay in high throughput. Left, example showing an *E. coli* plate and Iris simultaneous quantification of colony size and biofilm formation. Right, replicate reproducibility of *E. coli* biofilm formation.

The assay is based on a dye combination (congo red and coomassie blue, see 6.1.3.4), which binds to exopolysaccharides, and curli fimbriae produced by bacterial cells during biofilm formation (Barnhart and Chapman, 2006; Ghafoor et al., 2011; Van Houdt and Michiels, 2005). Banzhaf et al. and Herrera et al. independently applied the biofilm assay to existing mutant libraries of *P. aeruginosa*, and *E. coli* (Baba et al., 2006; Liberati et al., 2006). I developed the Iris software to quantify biofilm formation and found the assay to be reproducible

(r=0.96, see **Figure 7**) and largely independent of growth fitness for both species (see **Figure 8**). To establish that the assay also captures known pathways involved in biofilm formation in the two organisms, we measured the biofilm ability of each mutant in biofilm-inducing conditions.

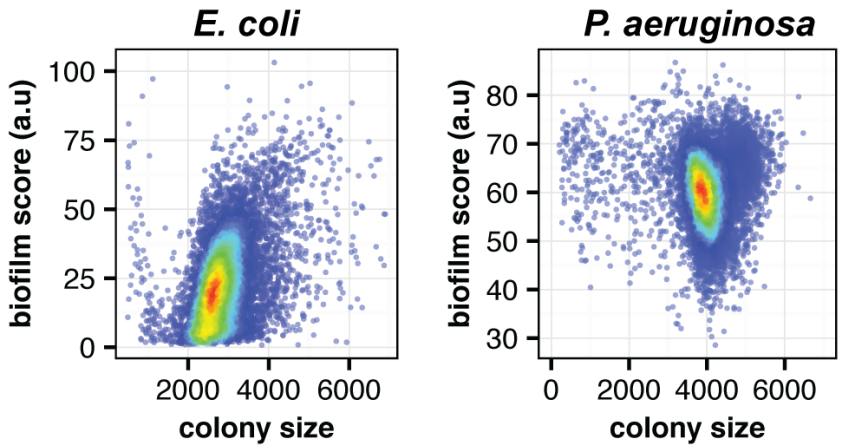

Figure 8: Biofilm quantification is largely independent of colony size. Left, biofilm and size data from colonies of genome-wide *E. coli* K-12 deletion mutants (Baba et al., 2006). Right, biofilm and size data from colonies of *P. aeruginosa* PA14 transposon mutant library (Liberati et al., 2006).

In the case of *E. coli*, I found that lipopolysaccharide biosynthesis is enriched among mutants with decreased biofilm formation. As expected, mutants of the curli fimbriae synthesis pathway are also unable to form biofilm with the exception of *ΔcsgC* (see **Figure 8**B). CsgC is a periplasmic protein that was recently shown to prohibit the major curli fiber subunit CsgA from forming intracellular amyloid fibers (Evans et al., 2015). At the same time, a *ΔcsgC* strain was shown to have little effect on congo red binding compared to the WT strain, as verified in our screen.

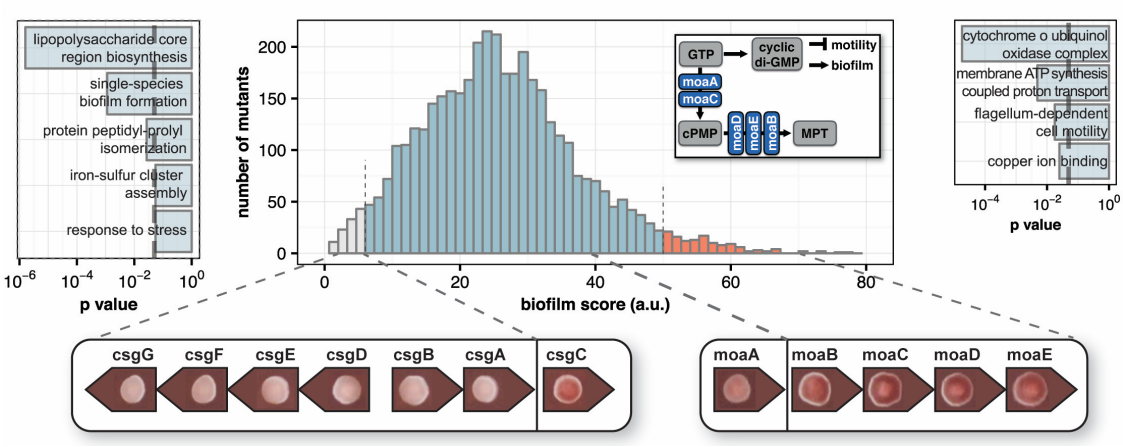

Figure 9: *E. coli* KEIO collection (Baba et al., 2006) biofilm phenotype distribution and outlier thresholds (2.5% each side); left, right: outlier GO enrichments of outlier mutants (BH-corrected p-values, dotted line is p-value 0.05). Examples demonstrate the LPS biosynthesis mutants deficiency in biofilm formation, and the Mo-molybdopterin biosynthetic process showing increased biofilm formation.

Strikingly, mutants of the molybdopterin (MPT) biosynthesis pathway are found to have highly increased biofilm formation. One possible explanation is that both MPT and cyclic-di-GMP biosynthesis processes draw from the same pool of GTP. Thus failure to synthesize MPT could lead to elevated levels of cyclic-di-GMP (see below), and consequently decreased motility and increased biofilm (Hengge, 2009).

We also screened the *P. aeruginosa* PA14 transposon (Tn) mutant library (Liberati et al., 2006) in a similar fashion to the *E. coli* library (see Materials and Methods). Among the complex signals that govern biofilm formation in *P. aeruginosa*, cyclic-di-GMP is among those that are in common with *E. coli* and other bacteria (Hengge, 2009; Jenal and Malone, 2006; Simm et al., 2004). Cyclic-di-GMP is synthesized from GTP by diguanylate cyclases (DGCs) and degraded by phosphodiesterases (PDEs). DGCs feature a GGDEF catalytic domain and DGCs feature either an EAL or a HD-GYP catalytic domain. I found that a number of DGCs mutants, in addition to the known *wspR* (Guvener and Harwood, 2007) and *sadC* (Merritt et al., 2007; Moscoso et al., 2014), have decreased biofilm formation (see Figure 10). Some genes carry both GGDEF and EAL domains on the same polypeptide, but often only one of the 2 domains is active with most known cases acting as hydrolases (Hengge, 2009; Jenal and Malone, 2006). In contrast to our expectation, all these mutants exhibited lower biofilm formation. A closer inspection of the transposon insertion site in these mutants revealed that in all cases it did not disrupt the EAL domain and could

potentially have led to its overexpression or mis-regulation, which could in turn explain the observed effects.

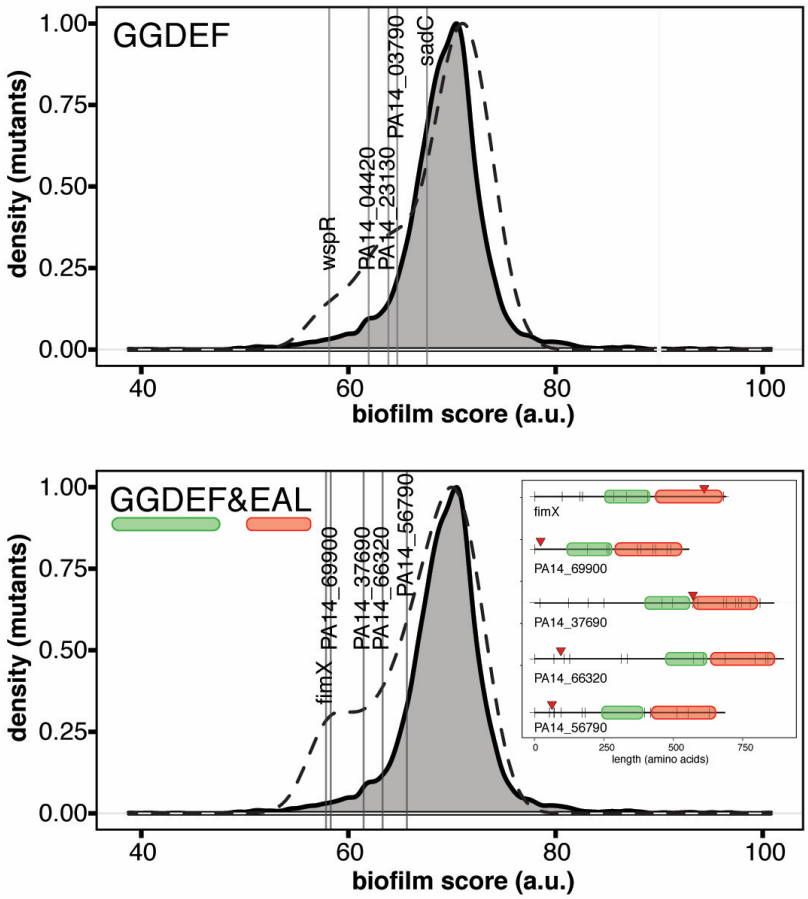

Figure 10: *P. aeruginosa* PA14 Tn library biofilm phenotype distribution. Overlay density plots (dotted lines) demonstrate the phenotypes of mutants in genes annotated with a GGDEF (upper plot), or both GGDEF and EAL domains (lower plot).

A special case is FimX, which also holds both GGDEF and EAL domains. Its GGDEF domain is catalytically inactive, however it has been shown to bind cyclic-di-GMP, thereby allostericaly regulating FimX function. So FimX works as a cyclic-di-GMP effector rather than as an enzyme (Jain et al., 2012; Navarro et al., 2009; Qi et al., 2011). Moreover, FimX has a long-established role in the formation of Type IV pili, which are surface structures implicated in twitching motility, adherence, and biofilm formation (Huang et al., 2003).

As expected, a mutation in any of the Type IV pili machinery and chemotaxis genes resulted also in decrease in biofilm formation (see Figure 11). Sole exceptions were mutants of genes *pilW*, *pilC* and *fimU*. These genes all harbor Tn

insertions at the first few nucleotides of their sequence, which likely results in their overexpression or deregulation rather than their deactivation (Liberati et al., 2006).

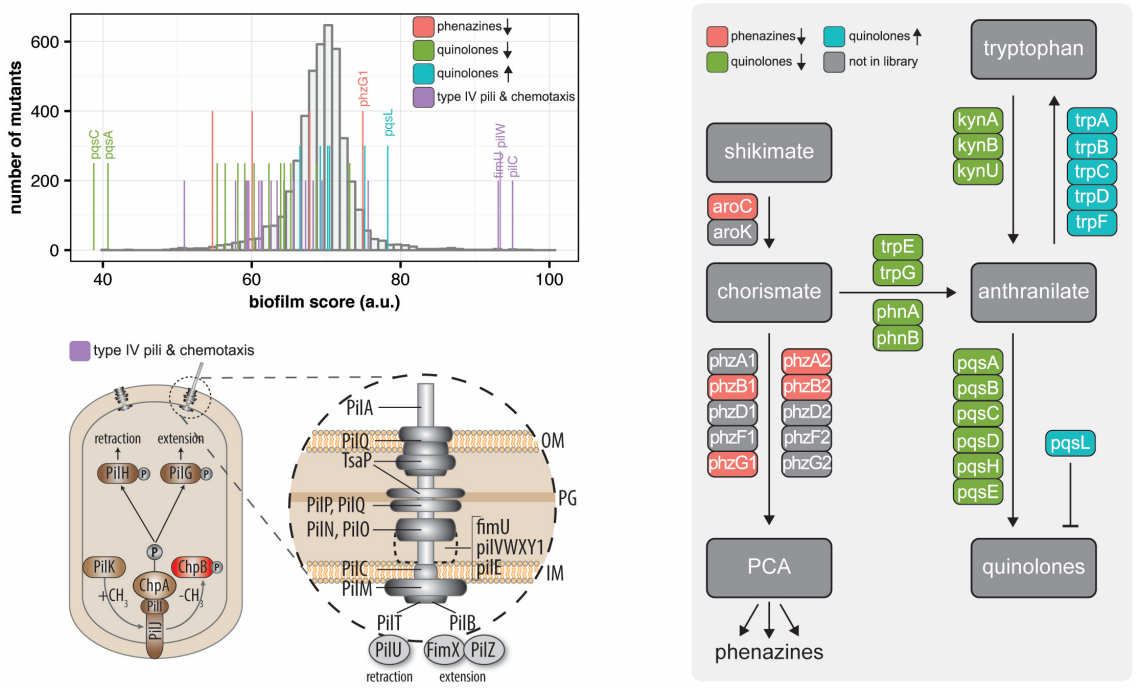

Figure 11: *P. aeruginosa* PA14 Tn library biofilm phenotype distribution. Density plots demonstrate the phenotype of selected mutant groups: purple (bottom left) twitching motility mutants, including mutants of typeIV pili machinery and chemotaxis show severely impaired ability to form biofilms (exceptions are PilC and 2 of the minor pilins: FimU and PilW). Right: mutant groups in the phenazine (red) and quinolone (green) biosynthesis pathways also show decreased biofilm formation; *pqsL* mutant shows increased biofilm formation

I also focused on the role of quorum-sensing pathways on *P. aeruginosa* biofilms. I examined the quinolone and phenazine production pathway, both of which serve as important quorum sensing (QS) molecules. Phenazines are a quite diverse set of compounds, which have been implicated in electron transport, iron uptake, and signaling (Pierson and Pierson, 2010; Wang et al., 2011). Of special interest to *P. aeruginosa* is the phenazine pyocyanin, which has been shown to promote biofilm formation (Dietrich et al., 2006). Phenazine production pathway branches after phenazine-1-carboxylic acid (PCA), a common phenazine, which was also shown to increase biofilm formation in an iron-dependent manner, by reducing ferric iron to ferrous iron (Wang et al., 2011). As expected, I observed that a mutational block of the PCA biosynthesis pathway resulted in decreased biofilm formation. I also observed a similar phenotype in mutants impaired in

quinolone biosynthesis. *P. aeruginosa* is known to use the quinolone PQS and its precursor HHQ as signals involved in virulence and biofilm formation (Nadal Jimenez et al., 2012). On the other hand, a *pqsL* mutant was shown to result in increased PQS production (D'Argenio et al., 2002), which expectedly results in highly increased dye staining in our assay. Mutants of genes involved in processing anthranilate, a quinolone precursor funneling it into or away from quinolone synthesis had a consistent negative or positive effect respectively on biofilm formation.

Notably, I found that the assay also reports interesting links to central metabolism. In contrast to previous biofilm studies in liquid (Musken et al., 2010), I found that several mutants involved in L-arginine biosynthesis show elevated biofilm production (see Figure 12 left) when grown on agar. To verify these observations, I queried mutants of this pathway after supplementing with amino-acid products of each intermediate step of this pathway. Using Iris I was able to quantitatively assess these results and noticed that addition of arginine slightly but reproducibly affected the pigmentation of the WT. At the same time, this addition alleviated the increased biofilm formation of the *argG* and *argH* mutants, which code for the last steps in the L-arginine biosynthesis pathway (see Figure 12 right).

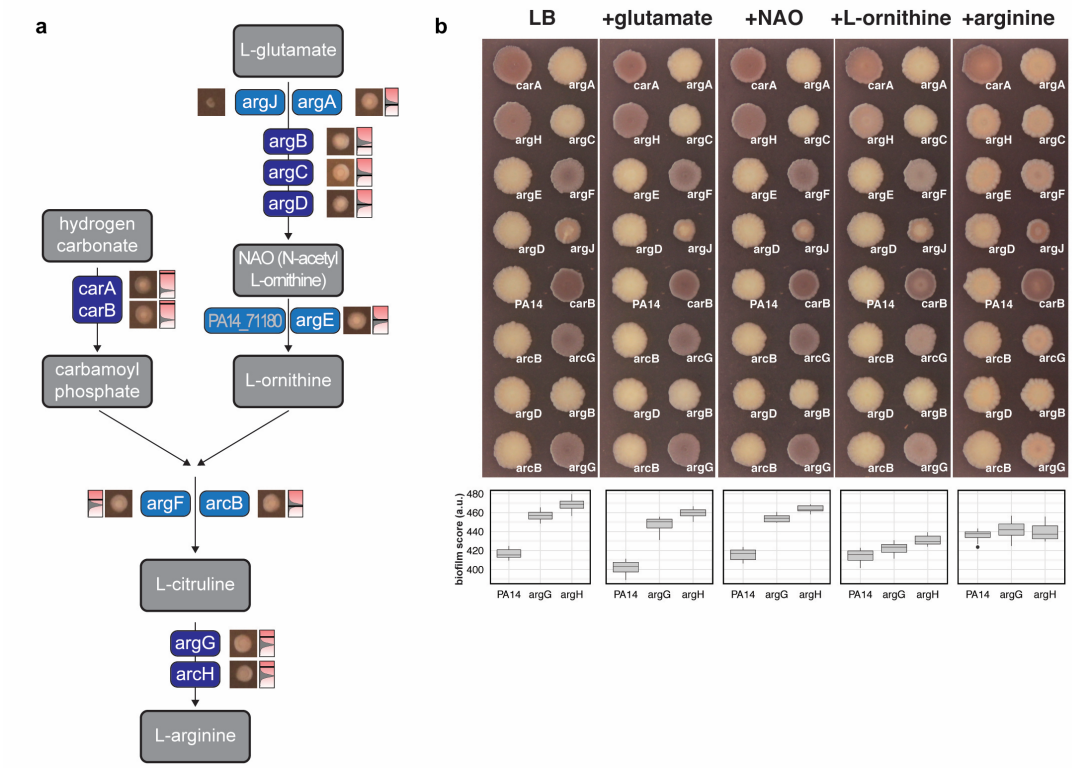

Figure 12: Biofilm biosynthesis phenotypes of *P. aeruginosa* PA14 mutants in arginine biosynthesis pathway (via L-ornithine). a. Measured biofilm phenotypes of each mutant is denoted by a black line. Background distribution corresponds to biofilm scores of all PA14 mutants grown on LB b. Biofilm formation phenotypes of arginine biosynthesis mutants grown on LB media supplemented with arginine biosynthesis pathway intermediate compounds. Iris-quantified biofilm phenotypes shown for the WT (PA14), and the *argG* and *argH* transposon mutants.

My contribution in both projects was to extend Iris to perform the biofilm readout. Development of the software was done hand-in-hand with the optimization of these high-throughput assays, to minimize detection issues and improve biofilm readout dynamic range. Subsequently, I performed a large part of the data analysis for each of these screens, especially with regard to detecting and correcting for biofilm readout technical biases.

Both of these screens have yielded a plethora of exciting results, including phenotypes for up to 85% of the gene perturbations in each assayed species. Results provide new leads on biofilm biology, as well as condition-specific regulation of biofilm development. Follow-up on such leads is currently being completed, and two independent manuscripts in which I will be a co-author are expected in the near future.

2.4.4 Colony morphology

Microbial colonies often form biofilms that feature higher order morphological structures. Colony morphology is a macroscopic observation of the behavior of cells growing in the biofilm. For instance in the opportunistic pathogen *Candida albicans*, colony structure complexity has been linked to the growth mode of cells (Ryan et al., 2012). *C. albicans* can either grow as spherical cells or form hyphae. This switch in cell appearance has been linked to virulence, as well as immune system evasion strategies (Bastidas and Heitman, 2009; Mitchell, 1998).

Moreover, in bacterial species such as *E. coli*, *P. aeruginosa*, and *B. subtilis*, colony morphology has been extensively studied with respect to cell reaction under nutrient limitation occurring within the biofilm. Specifically colony morphology studies in *P. aeruginosa* showed that colony wrinkling promotes oxygen access to cells growing in biofilm (Madsen et al., 2015). Interestingly, *P. aeruginosa* strains that commonly appear in chronic human lung infections have been shown to have hyper-wrinkled colony morphologies (Starkey et al., 2009). In *E. coli* and *Salmonella enterica*, morphology has been shown to depend on flagella function, as well as cellulose and curli fimbriae production (Prigent-Combaret et al., 2000). Recent studies employed scanning electron microscopy to reveal a highly ordered localisation of cellulose and curli fibres within *E. coli* colonies (Serra et al., 2013).

Colony morphology is one of the elementary steps of characterizing a microbial colony. However, this characterization has been a descriptive, qualitative observation, even when it has been applied at a larger-scale level (Cabeen et al., 2016; Ryan et al., 2012). Iris can quantify colony structure complexity by using a novel colony morphology detection and quantification algorithm. I demonstrate that the algorithm is capable of quantifying microbial colony morphology across several diverse species. This enables us to identify morphology regulators and generate new hypotheses.

Colony morphology quantification complements other phenotypes Iris can quantify, adding another measurement that can be used to analyze bacterial colonies. I demonstrate that combining the colorimetric biofilm readout with the morphology readout can lead to dissecting the biofilm biosynthesis pathway in the model species *Salmonella enterica* serovar Typhimurium.

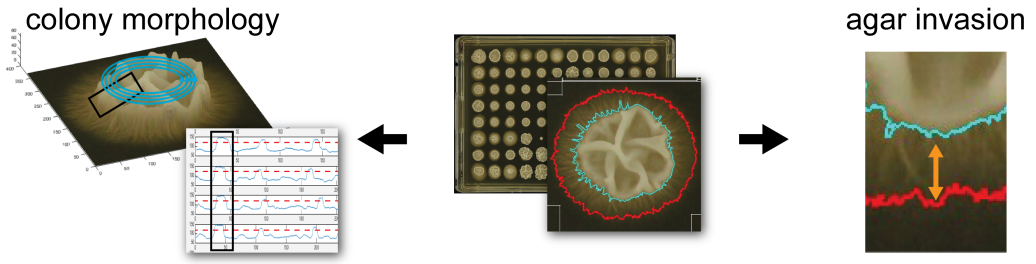

Figure 13: Colony morphology quantification in *Candida albicans* colonies. Left, an illustration of the morphology complexity quantification algorithm, colony is traversed in concentric circles (light blue), brightness values are visualized as in the inlet, and brightness peaks higher than the threshold (red dotted line) are counted. Right, measuring the extent of colony agar invasion is done using two thresholding algorithms of different sensitivity, boundaries of in-agar growth are shown in red, while over-agar growth boundaries are shown in blue.

Since existing texture-detection methods could not account for technical aspects, such as lighting differences (Howarth and Rüger, 2004), I developed a new colony structure detection algorithm. For a description of the colony morphology quantification algorithm, I direct the reader to the Materials and Methods section 6.1.3.5. Briefly, the algorithm traverses colony pixels in concentric circles to detect number of ridges in brightness values (see **Figure 13**).

2.4.4.1 Candida albicans

Iris was used to quantify the structure complexity and agar invasion of *C. albicans* colonies of two homozygous single-gene deletion collections (Homann et al., 2009; Noble et al., 2010). The human pathobiont fungus *C. albicans* can grow *in vivo* as yeast, hypha or pseudohypha, a morphogenic switch that has been linked to virulence (Sudbery et al., 2004). Phenotypically, *C. albicans* colony structure on solid media reflects the three cell types. A smooth colony comprises mostly of yeast cells, while a wrinkled one comprises mostly of hyphae and pseudohyphae (Ryan et al., 2012). Invasive filamentation occurs around the colony as hyphae and pseudohyphae penetrate the agar. Here, I used Iris to simultaneously score the colony structure and invasive filamentation of homozygous deletions of 674 genes (11% of the genome, (Noble et al., 2010)) and of 143 transcriptional factors (Homann et al., 2009). The mutants were arrayed on agar plates with Spider medium and incubated at 30C for 7 days. The algorithm successfully rated deletions of genes that are part of a regulatory network of biofilm formation and known to be crucial for filamentation (Fox et al., 2015), linked to biofilm formation (Nobile et al., 2012; Schweizer et al., 2000; Sellam et

al., 2010), and morphogenesis changes upon several stimuli, such as pH changes (Davis et al., 2000; Kim et al., 2008; Stoldt et al., 1997) (see Figure 14).

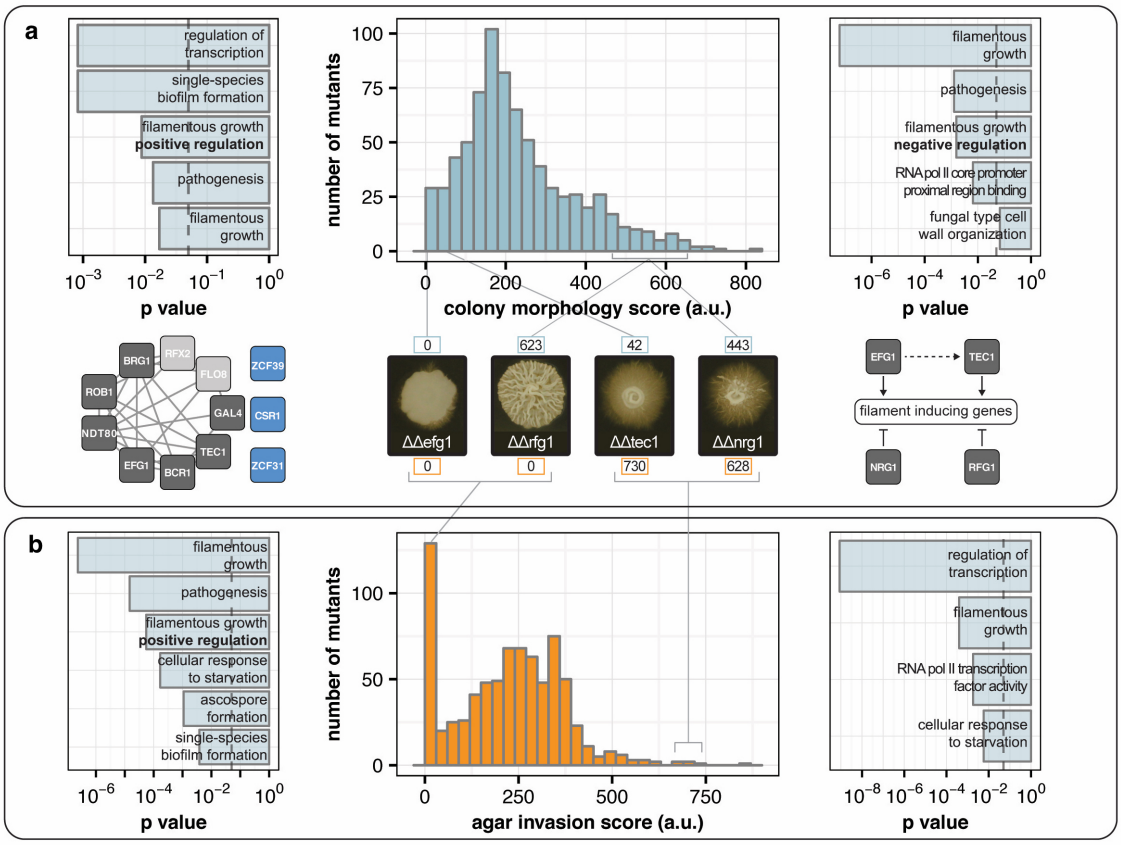

Figure 14: Quantitative measurements of *C. albicans* mutants colony morphology and agar invasion. (a) *C. albicans* mutant collection colony morphology phenotype distribution and outlier GO enrichments of outlier mutants (BH-corrected p-values, dotted line is p-value 0.05). Examples of morphology-impaired outliers include a network of transcription regulators that mediate biofilm formation, while examples of mutants showing increased colony structure complexity include repressors of filamentous growth Nrg1 and Rfg1. Colony pictures of mutants from the colony arrays and demonstrate that readouts are orthogonal, e.g. that morphology outliers (Iris-reported value over colony picture) do not necessarily show increased in-agar growth (Iris-reported value under colony picture) and vice-versa. (b) *C. albicans* mutant collection in-agar growth phenotype distribution and outlier GO enrichments of outlier mutants (BH-corrected p-values, dotted line is p-value 0.05).

Iris was also able to measure invasive filamentation, genetic determinants of which are shown to be distinct from those of colony morphology. For example deletions of *ccn1*, a G1 cyclin (Loeb et al., 1999), *rfg1*, a transcriptional repressor of filamentation (Khalaf and Zitomer, 2001), *ppg1*, a protein phosphatase (Albataineh et al., 2014) or *pep7*, a vesicle transport protein (Franke et al., 2006) don't invade the agar, even though their colony structures differ. On the opposite end of the spectrum, mutants *ngr1*, a transcriptional repressor of filamentation

(Murad et al., 2001) and *rbd1*, a rhomboid-like protein (Noble and Johnson, 2005) are quantified to form colonies with higher in-agar filamentation (see Figure 14).

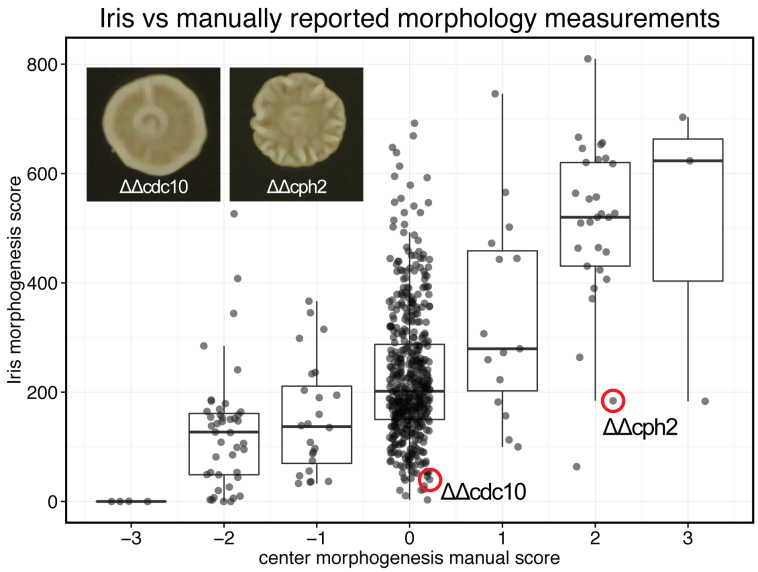

Figure 15: Iris-reported morphology values for *C. albicans* mutans compared to expert user manual annoatation. Two highlighted mutants Iris reports above average, while manual annotation reported as average or above average correspond to genes involved in hyphal growth.

Colony morphology values and expert user manual morphology scores were then compared and found to correlate well (see **Figure 15**). Here I show examples of two cases where Iris correctly quantified mutant colonies as less structure-forming than the expert user. Both such cases involve genes related to hyphal growth. Cph2 is a transcription factor that promotes hyphal growth (Lane et al., 2001), while Cdc10 is a septin required for normal hyphal growth.

2.4.4.2 Salmonella enterica

Colonies of *Salmonella enterica* and related species are often described by their biofilm production and morphological features. Congo-red agar plates have been a long established assay to study biofilm. Wild type *Salmonella enterica* develops red dry and rough (rdar) colonies on such plates. Pathways that majorly affect this phenotype are cellulose and curli fimbriae biosynthesis (Zogaj et al., 2003; Zogaj et al., 2001).

With the help of Lucia Herrera (Typas lab), I screened and analyzed all genes in a *S. enterica* serovar Typhimurium mutant library (Porwollik et al., 2014) simultaneously for colony color and structure complexity. I show that Iris can accurately quantify both of these phenotypes of *S. enterica*. Moreover, I show that the combined data can be used to dissect the cellulose and curli fimbriae biosynthesis pathways (see **Figure 16**), both of which serve as biofilm components.

As expected, a block in cellulose production results in colonies with no structures. At the same time, deleting *csgD* the gene coding for the central regulator of biofilm formation (Simm et al., 2014), or any of the genes controlling this regulator results in loss of both components of the biofilm, thus in colonies with severely reduced morphology and color (see **Figure 16**).

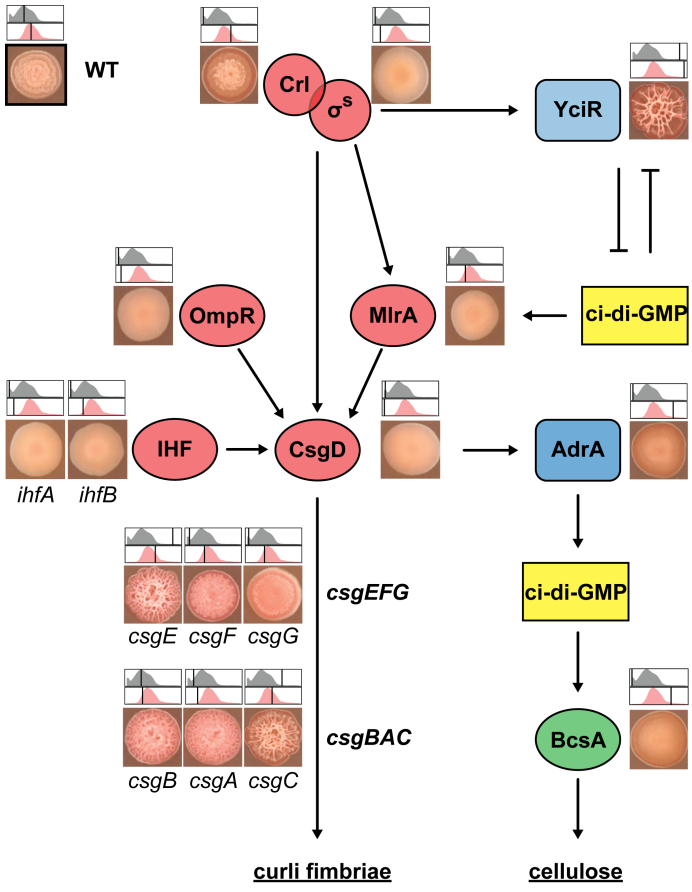

Figure 16: A comprehensive library of *Salmonella enterica* serovar Typhimurium single gene deletion mutants (Porwollik et al., 2014) was assayed simultaneously for Congo red binding and colony structure formation (colony morphology). Histograms compare the quantitative measure of Congo red binding (red) and the colony morphology (gray) of all mutants in the library versus the values of shown mutants (black line). Mutants in curli fimbriae and cellulose synthesis pathways show distinct phenotypes, for instance a mutational block in cellulose synthesis abolishes colony morphology but retains a more red colony color, implying compensation from the curli fimbriae pathway.

2.4.4.3 *Pseudomonas aeruginosa*

A transposon mutant library of *P. aeruginosa* PA14 (Liberati et al., 2006) was assayed for biofilm formation and colony morphology. By acquiring pictures at different timepoints, I was able to independently select timepoints exhibiting maximum dynamic range for the quantification of both color development and colony structure complexity (see **Figure 17**).

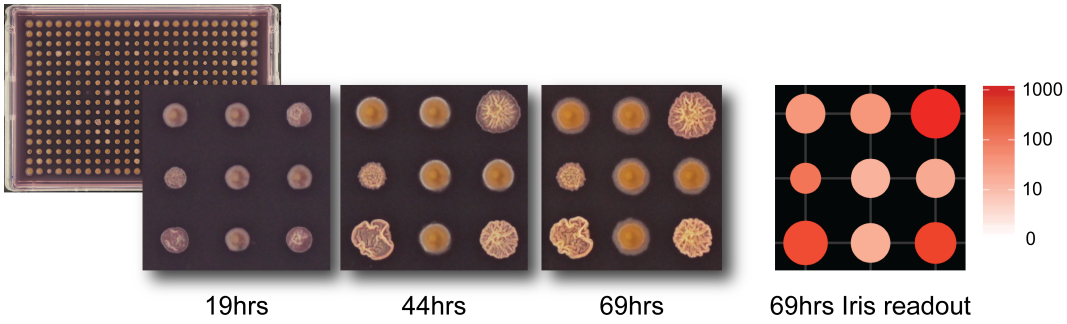

Figure 17: Colony morphology quantification in colonies of *Pseudomonas aeruginosa* across timepoints. Right, example of Iris quantification for the 69 hours post-inoculation.

By combining the quantitative measurements of those 2 Iris readouts, I generated a 2-D map of all *P. aeruginosa* PA14 mutants with respect to their biofilm formation and colony structure complexity. Moreover, I placed on this map the phenotypes of mutants known to be involved in flagella, type IV pili apparatus and regulation, and mutants related to the signaling pathways involved in quorum sensing.

Interestingly, this phenotypic approach indicates that mutants of a gene of unknown function (PA14_14210), as well as the homologue of dipeptide transporter protein DppD result in highly increased colony structure formation (see **Figure 18**). In fact, such high levels of colony structure formation are only observable in mutants of the flagella apparatus thus forming the foundation for an interesting hypothesis, which we are currently following up in the lab.

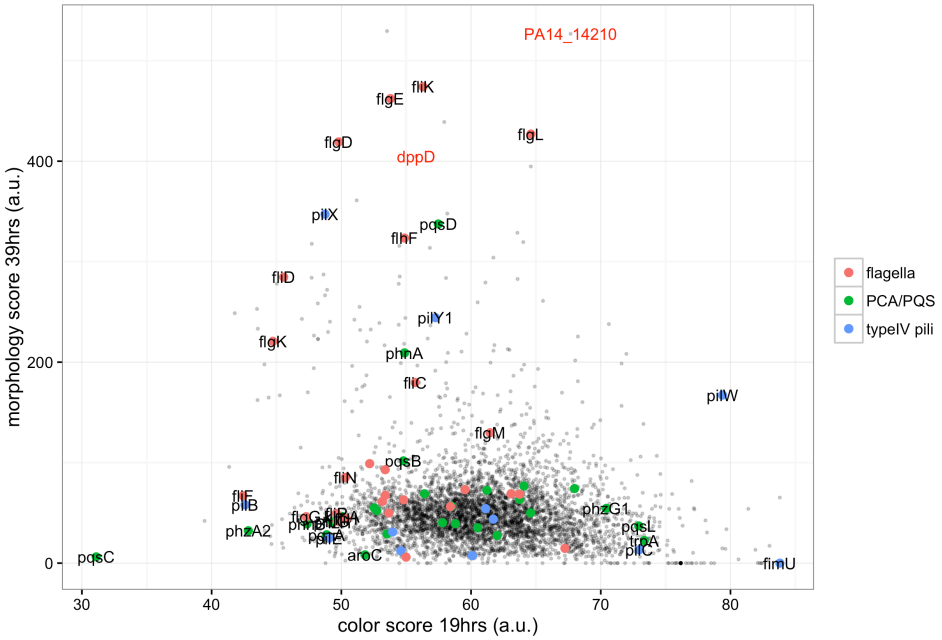

Figure 18: Congo red binding and colony morphology formation compared across colonies of a *P. aeruginosa* PA14 mutant library (Liberati et al., 2006). PCA/PQS: Phenazine or quinolone biosynthesis mutant.

2.5 Outlook

I have developed an image analysis tool called Iris that can be used quantify several microbial colony phenotypes in high throughput. Iris is freely available, and already in use by several labs across the globe. Moreover, Iris is open-source so that future developers can modify it to the needs of their lab.

Iris is designed for easy expansion on new readouts. In the near future I will be expanding the software to accommodate the readout of diverse phenotypes exhibited by natural isolates of *E. coli*. In this case, I combined Iris readouts with a machine learning approach to discern classes of different microbial phenotypes. I expect this approach to soon give fruit, which could also feed back to Iris. For example, making a better decision on colony phenotype class, could direct Iris to use the best colony detection module for this class.

This machine learning approach is already underway in an application of detecting microbial species interactions. Indeed one can query a microbial library next to a species of interest by arraying them side-by-side on an agar plate (see **Figure 19**). Iris can then independently quantify the phenotypes of both assayed species, resulting in a quick, unbiased way of reporting phenotypes of cross-

species interactions. Importantly, assaying mutant libraries of either species can quickly result in leads towards the molecular basis of inter-species interaction.

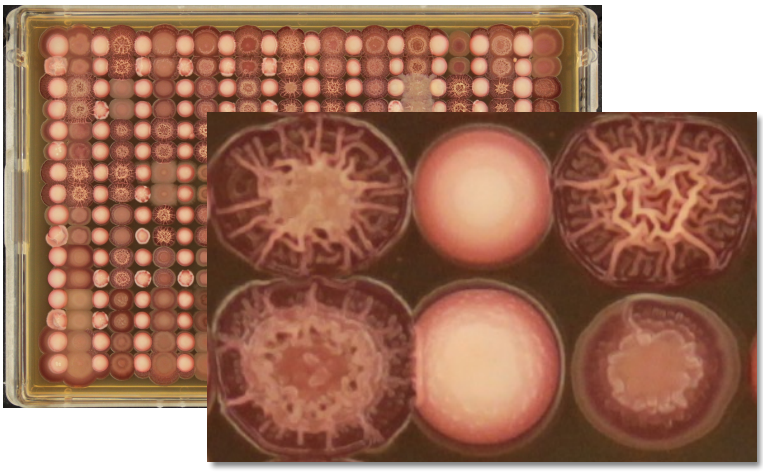

Figure 19: Cross-kingdom inter-species interactions. Colonies of *E. coli* single gene deletion mutants (inlet corners) exhibit highly variable phenotypes when arrayed side by side *C. albicans* colonies (inlet middle).

Finally, for the near future, I plan to adapt Iris to also perform microbial colony phenotype characterization in a low-throughput setting. Iris is already successfully used in several challenging high-throughput applications across several labs. However, several more labs exist without high-throughput needs or infrastructure. By adapting Iris to be used for low throughput single-colony readout I expect it to become a useful quantitative phenotyping tool in many more microbiology labs.

2.6 Contributions

Pseudomonas aeruginosa experiments (both biofilm and morphology assays) were performed by Manuel Banzhaf (Typas lab, EMBL). *Salmonella enterica* experiments were performed by Lucia Herrera (Typas lab, EMBL). *Bacillus subtilis* ordered mutant library preparation and assays were performed by Byoung-Mo Koo (Gross lab, UCSF). Experiments involving *Candida albicans* and cross-kingdom species interactions were performed by Alexandra Koumoutsi (Typas lab, EMBL). Kinetics data for growth of *E. coli* colonies were acquired by Anja

Telzerow (Typas lab, EMBL) and myself. CPRG-agar plate experiments were performed by Nassos Typas, Thomas Bernhardt, and Catherine Paradis-Bleau. For all the above, I developed the software to quantify these diverse phenotypes, and subsequently analyzed the phenotypic data. The small-scale *E. coli* chemical genomics screen on neglected antibiotics was performed by Anthony Shiver (Gross lab, UCSF), data were analyzed by Anthony Shiver and myself.

3 Uncovering genetic determinants of envelope biogenesis in Gram-negative bacteria

3.1 Background and significance

The bacterial cell envelope is the first line of defense against a multitude of environmental challenges. In Gram-negative bacteria, the cell envelope is consists of a rigid peptidoglycan sacculus sandwiched between two lipid bilayers, named the inner and the outer membrane. The peptidoglycan cell wall gives bacterial cells the ability to maintain their shape, and to withstand osmolarity changes. In addition, incorrect synthesis or maintenance of this rigid cell wall often leads to cell death as a result of their high turgor pressure.

On the other hand, the outer membrane (OM) layer of Gram-negative bacteria acts as a molecular sieve, restricting uncontrolled entry to large, or hydrophobic compounds. The OM is crossed by porins that act as diffusion channels, while the OM outer leaflet consists mainly of charged lipopolysaccharide (LPS), which makes the cell impermeable to hydrophobic molecules. Importantly, changes to the LPS and the porins are involved in pathogen survival during infection, as well as antibiotic resistance.

The asymmetric nature of the OM lipid bilayer makes Gram-negative bacteria more impermeable to antimicrobial compounds. Both hydrophobic compounds, as well as large hydrophilic compounds are excluded from entering. This fact combined with an armory of efflux pumps makes these organisms notoriously hard to target with antibiotics. The situation is further aggravated by the rise in multi-drug resistant microbial strains in clinical settings (Davies and Davies, 2010; Fernandez and Hancock, 2012).

Many such strains have ways to reduce the effective concentration of the drug inside the cell, such as modifications to their cell envelope, or adaptation of the drug target. Therefore, new treatments against these organisms are urgently needed. Such a goal will be greatly aided by a detailed mechanistic understanding of cell envelope assembly and maintenance. While our knowledge of this process

has advanced in the past years, many aspects of cell envelope biogenesis still remain to be elucidated.

3.2 Results

3.2.1 Bacterial envelope permeability assay

With the goal of identifying factors required for correct cell envelope biosynthesis, we developed a simple screening process involving the reporter enzyme β-galactosidase. The classic β-galactosidase enzyme activity assay is based on measuring the processing of ortho-nitrophenyl-b-D-galactopyranoside (ONPG) into the yellow ortho-nitrophenol (ONP). Importantly, this assay involves lysing the bacterial cell membrane, to allow for substrate contact with the intracellular enzyme (Miller, 1972).

In contrast, intact cells do not allow the substrate to enter the cell. Thus only mutants defective in membrane biosynthesis or with elevated lysis levels will allow for the reaction to occur, which leads to their detection. As a substrate, we chose chlorophenol red-beta-D-galactopyranoside (CPRG), which was also shown to be more sensitive than the ONPG-based assay (Eustice et al., 1991). CPRG gets processed by β-galactosidase into the red chromophore CPR, which allows for easier detection of the reaction product on agar plates compared to the yellow ONP product (see **Figure 20**).

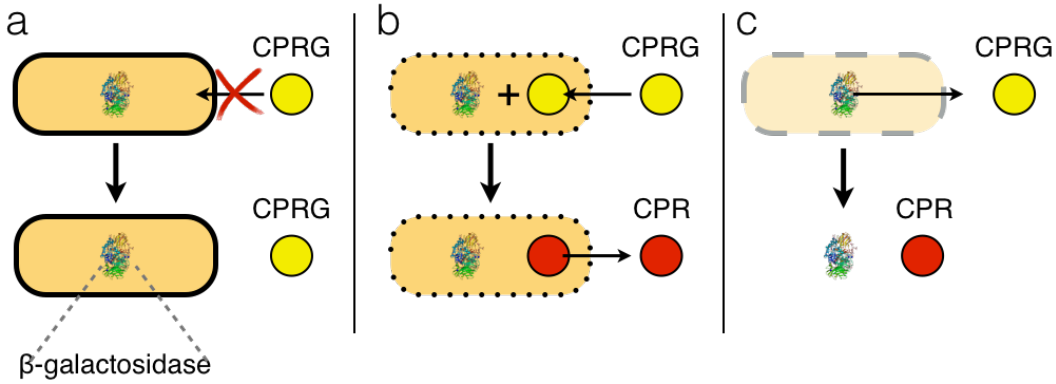

Figure 20: Permeability assay schematic. (a) Wild-type cells are unable to cleave CPRG, because the enzyme (β-galactosidase) is separated from the enzyme by the intact cell envelope. (b) Cells with impaired envelope function are permeable to CPRG, which is processed by β-galactosidase into the red chromophore CPR. (c) Cells that lyse release the β-galactosidase enzyme into the medium where it can freely process CPRG into CPR.

3.2.2 Image-based high throughput screen for mutants defective in envelope biogenesis

To acquire an unbiased genome-wide view of envelope biogenesis mutants, we screened an ordered *E. coli* mutant library (Nichols et al., 2011) that includes the KEIO collection (Baba et al., 2006), as well as a collection of mutants with hypomorphic alleles of essential genes and mutants lacking genes for small RNAs.

To allow rapid phenotype acquisition, mutant colonies were arrayed on 384-format plates. After incubation in 4 different conditions, CPRG turnover was measured using automated image analysis software.

While key results of this study are outlined in this paragraph, I direct the reader to the manuscript published in PLoS Genetics (Paradis-Bleau et al., 2014).

3.2.2.1 Sensitive detection of colony hue alteration

In permeable bacterial cells, or cells that undergo lysis, the intracellular β-galactosidase will process CPRG into the red chromophore CPR. Thus colonies of such mutants arrayed on solid agar surface will turn red. I developed the Iris image analysis software to quantify a range of bacterial colony phenotypes, and subsequently implemented a readout specialized to detect colonies with different pigmentation.

Owing to diffusion, one of the limitations of the CPRG readout on agar plates was that accurate readout could only be performed in an early timepoint. I thus developed a sensitive method to detect hue changes, by converting colony pictures to the HSV color space. For a more detailed description of the image analysis and phenotype quantification process, I direct the reader to the chapter related to the Iris software (see page 7).

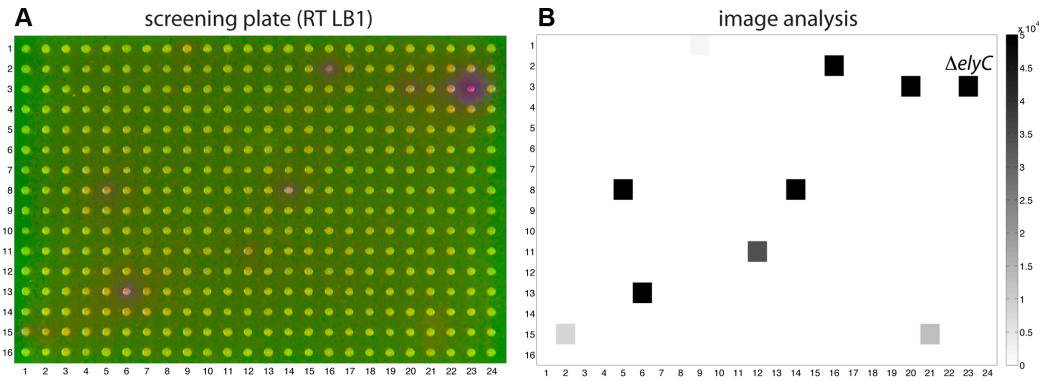

Figure 21: Example of Iris quantification of a developed CPRG plate. ΔelyC strain can be seen on the top right. A. CPRG plate inoculated with one of the mutant library plates in array format and developed for 23 hours at room temperature. B. Quantification of the same plate based on image analysis software Iris. Figure is adapted from (Paradis-Bleau et al., 2014).

3.2.2.2 Screen reveals numerous genes with novel phenotypes

In order to maximize the number of identified factors responsible for envelope biogenesis, we decided to screen the ordered mutant library at room temperature and 30°C, on media with different NaCl concentrations (0 and 1% NaCl for both temperatures). This approach can also potentially identify factors with a role in adaptation to different temperatures or osmotic conditions.

After incubation in different conditions and automated detection of CPR signal for all colonies, we set a CPR signal cutoff that resulted in few hundred CPR+ mutants in each condition. These CPR+ mutants often correspond to genes that are already implicated in correct cell envelope biogenesis. Indeed a GO enrichment analysis on these genes reveals a high enrichment for genes involved in lipid metabolism, LPS, and enterobacterial common antigen biosynthesis.

Calculating the overlap among hits of different conditions revealed a high number of genes that show a condition-specific CPR+ phenotype. At the same time, there are only a few genes responsible for a CPR+ phenotype across all conditions (see **Figure 22**B).

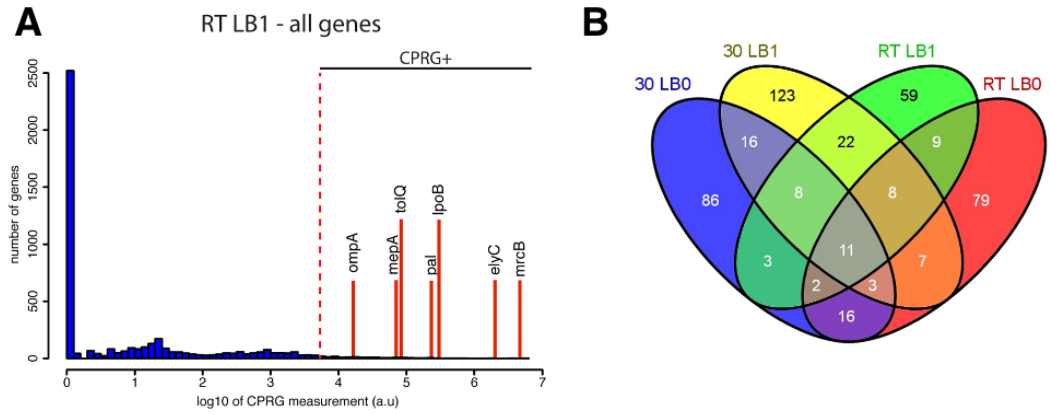

Figure 22: A. CPRG assay score distribution for the screen carried out at room temperature on agar prepared with 1% NaCl. Scores for genes of known importance for cell envelope integrity are indicated by the red lines. Genes with scores above the cut-off ($10^{3.7}$ units) were designated as CPR+ hits. B. Venn diagram comparing the CPR+ hits identified in the different growth conditions. RT is room temperature, while 30 is 30°C incubation; LB0 and LB1 correspond to 0% and 1% NaCl in the media. Figure is adapted from (Paradis-Bleau et al., 2014)

The screen identified several genes of unknown function as factors involved in envelope integrity. Across all 4 conditions, mutations in 102 genes of unknown function were found to result in a CPR+ phenotype, corresponding to more than 22% of all hits. Interestingly, mutants of over 70 of these genes were found to be unresponsive when measuring colony size as a proxy of fitness across more than 300 conditions (Nichols et al., 2011).

This reveals the potential of the CPRG assay in larger-scale chemical genomics studies. Since the assay can demonstrably capture information orthogonal to that present in growth fitness screens, results from a CPRG screen across conditions can complement growth information, resulting in a clearer view of the biology behind specific phenotypes. At the same time, there was only a small overlap of genetic factors responsible for CPR+ phenotype across different temperature and osmolarity conditions. Thus the imaging-based agar plate screen results indicate that screening across different conditions may reveal different biological aspects, see also 3.3.5.1.

As proof of principle, Catherine Paradis-Bleau (former Bernhard lab) further followed up one of the genes of unknown function, deletion of which causes a strong CPR+ phenotype. Gene *ycbC* (renamed to *elyC*) encodes for a protein with two transmembrane domains and a DUF218 domain with predicted periplasmic localization. This domain is also present in proteins SanA and YgjQ, both of which show no CPRG nor growth phenotype. DUF218 domains are present

throughout the bacterial kingdom, suggesting an enzymatic activity, however, the function of DUF218 is yet to be elucidated.

The growth and morphology of *ΔelyC* cells was monitored at room temperature. *ΔelyC* cells grow indistinguishably from the wild type in exponential phase. However, when the *ΔelyC* culture reached stationary phase, cells began lysing after formation of membrane blebs. This phenotype is similar to bacterial cell death phenotype following treatment with beta-lactam antibiotics, suggesting a role of *elyC* in peptidoglycan (PG) biosynthesis. Further PG synthesis measurements revealed that PG synthesis is blocked in a *ΔelyC* strain. In line with previous observations (Prats and de Pedro, 1989), cell growth can continue without PG synthesis for about one mass doubling.

To pinpoint the role of *elyC* in PG biosynthesis, candidate multi-copy plasmids (Saka et al., 2005) were introduced to the ΔelyC strain. Each such plasmid encodes for a factor of PG biosynthesis, including UppS and PBP1B. UppS is responsible for producing the lipid carrier Und-P, which is likely limiting for the synthesis of lipid-linked precursors (Barreteau et al., 2009), while PBP1B performs the final polymerization and crosslinking reactions (Sauvage et al., 2008). Interestingly, overproduction of UppS but not of PBP1B was able to rescue the CPR+ phenotype of *ΔelyC*, suggesting a role of *elyC* in the lipid carrier metabolism, and not in PG incorporation.

For a more detailed description of the follow-up experiments and discussion, we direct the reader to the published manuscript (Paradis-Bleau et al., 2014).

3.2.2.3 Limitations of chromophore reaction screening on agar plates

Despite its demonstrated potential to uncover phenotypes in mutants with elevated lysis or defects in envelope biogenesis, the CPRG assay in agar plates is limited by chromophore diffusion. This diffusion of the red chromophore CPR can spread from a CPR+ mutant colony to neighboring colonies, thus interfering with accurate measurement of these colonies. Moreover, at later timepoints, chromophore diffusion is so extensive that no colony on the plate is unaffected (see **Figure 23**), making it impossible to distinguish colony pigmentation.

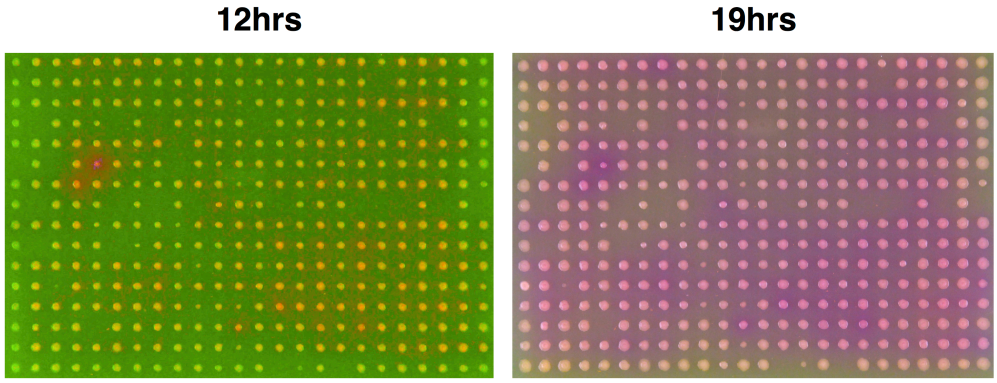

Figure 23: CPRG limitations of assay in agar plates. Example of an assay plate incubated at 30°C shows a high level of CPR chromophore diffusion by 19 hours post-inoculation. Diffused chromophore saturates the signal, making it impossible to distinguish CPR+ colonies.

This limitation inevitably results in a restricted development time for agar plates with mutant colony arrays. This restricted measurement time in turn results in capturing only the mutants that exhibit a CPR+ phenotype in the early colony growth phases. In fact, most mutants had no measurable phenotype despite sensitive chromophore detection in mutant colonies. This limitation effectively reduced the agar-plate CPRG assay to a qualitative assay, where only strong positive phenotypes can be accurately detected.

3.2.3 Prolonged quantitative measurement of envelope biogenesis defects

In order to overcome the main limitations of the agar plate membrane permeability assay, I built upon the assay and adapted it for automated, quantitative readout. The difference to solid agar plates of the preceding screening method is the use of agar-filled well plates. The use of well plates mitigates the main limitations of this assay, since there is no chromophore diffusion. We reasoned that colony growing on the surface of such an agar well will reach saturation faster than in a liquid setting, while still allowing for CPRG turnover in the agar medium.

This setup allowed for CPRG turnover measurements well into the stationary phase of the entire mutant library, accurately detecting several genes known to be implicated in correct cell envelope biogenesis. Moreover, by accurately measuring CPRG turnover over prolonged periods of time, I unprecedentedly uncover mutants that lyse less frequently than the wild type.

3.2.3.1 Disentangling CPRG turnover from colony growth

The CPRG to CPR turnover can also be measured by means of absorbance, which allows for the assay readout to be performed by a microplate reader. In such a setting, the plate reader light path traverses the agar-filled well as well as the colony growing on its surface. Moreover, the absorbance peak of CPRG is at 575nm, which is the same wavelength growth is measured in liquid. Thus, assuming constant well volume (see 6.2.3), the measured total absorbance will be influenced both by the colony growth, as well as by the CPR present in the medium.

Moreover, both growth fitness and CPRG turnover vary with time and genetic background. While the absorbance spectrum of most mutant colonies is practically level early in the experiment (see **Figure 24**A left), in the later timepoints several mutant colonies exhibit the characteristic peak at 575m to various degrees (see **Figure 24**A right). As an example shown in **Figure 24**B, the green line corresponds to the absorbance spectrum of the *ΔmrcB* mutant. *ΔmrcB* is shown also in the preceding agar plate screen to have a higher CPR turnover, which results in the higher absorbance peak at 575nm. As a negative control, mutant *ΔpcnB* (blue line) is known to have decreased number of plasmid copies (Liu and Parkinson, 1989) which results in reduced enzyme levels (see also 3.3.3).

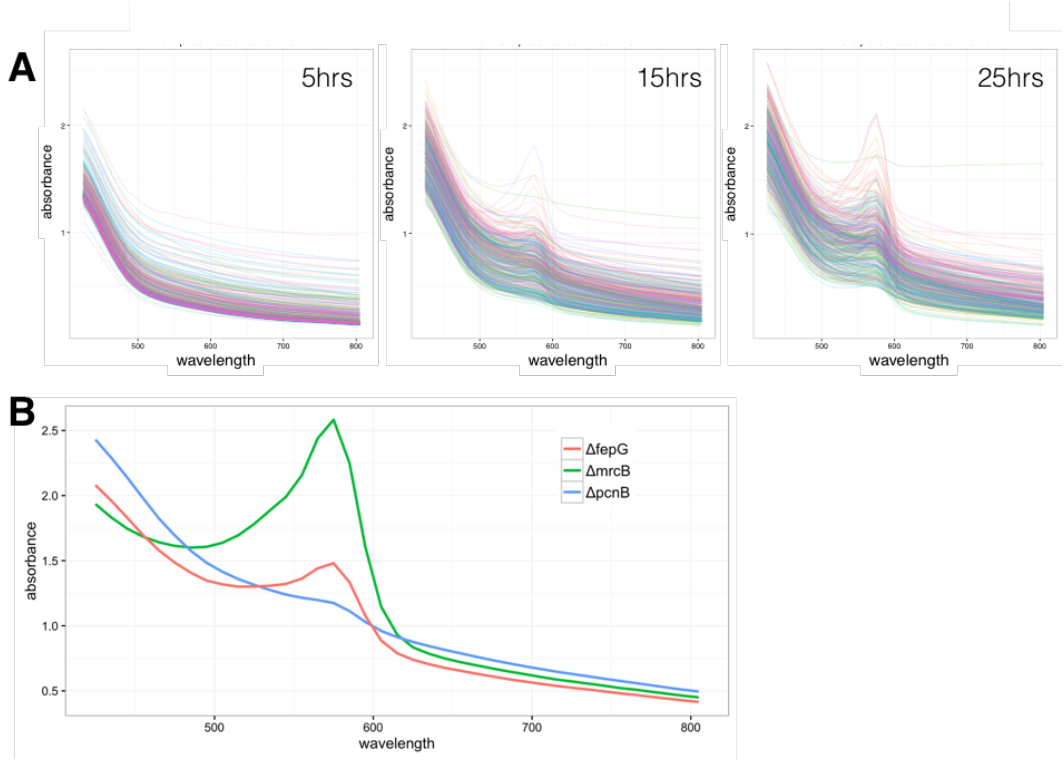

Figure 24: Spectral scans of *E. coli* colonies growing on agar-filled wells of a 384 microwell plate, CPR absorbance peak is at 575nm. A. Mutant colony spectral scans across different timepoints; CPRG turnover increases with time, as does the absorbance at 575 nm compared to the rest of the wavelengths. B. Spectral scans of mutant colonies of different genetic backgrounds at the same timepoint; *ΔmrcB* (green) has a higher absorbance peak at 575nm due to the increased CPR turnover, *ΔpcnB* (blue) has reduced plasmid copy number and thus reduced levels of β-galactosidase.

In order to avoid CPR chromophore measurements confounded by the colony absorbance, I developed a method to estimate the absorbance component introduced by the colony growth. With the help of Lucia Herrera (Typas lab) I acquired data on the absorbance spectra of 384 micro-wells with:

a) colonies of mutants that lack the lacZ gene, and thus cannot turn over CPRG into CPR
b) colonies of mutants that with an IPTG-induced lacZ plasmid, which also are permeable and do turn over CPRG into CPR
c) the purified β-galactosidase enzyme, which processes CPRG into CPR in the medium

I then observed that CPR chromophore has an absorbance peak at 575nm and absorbance minima at 450nm and 650nm, whereas colony growth exhibits a more level absorbance profile (see **Figure 25**).

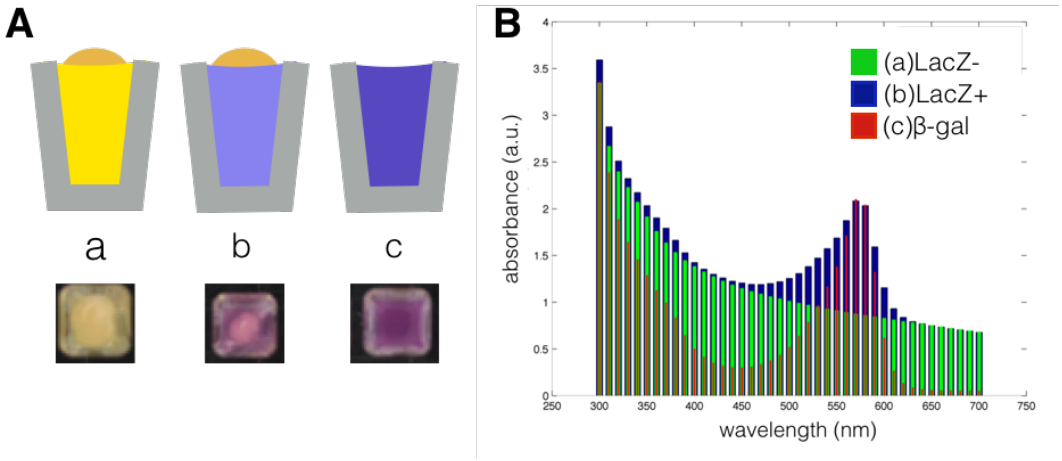

Figure 25: A. schematic representation and example pictures of (a) LacZ-, and (b) LacZ+ colonies growing on indicator agar well plates; (c) is a microwell filled with indicator agar and incubated with the purified β-galactosidase enzyme. Microwell plate reader light path traverses the well from top to bottom, hence absorbance is influenced by colony growth. B. Spectral scans of LacZ-, LacZ+ colonies and wells incubated with the β-galactosidase enzyme; last one (c) reveals absorbance minima of the CPR chromophore at 450 nm and 650 nm.

Subsequently, I trained a linear model to predict the absorbance at 575nm using as input absorbance measurements at 450nm and 650nm. The model was trained using data from several hundred colonies missing the lacZ gene, and thus in absence of the CPR chromophore. This model was shown to accurately predict the growth-related absorbance at 575nm with residuals ranging typically between ±0.02 (see **Figure 26**). This predicted growth related absorbance at 575 nm is subsequently used as a growth metric.

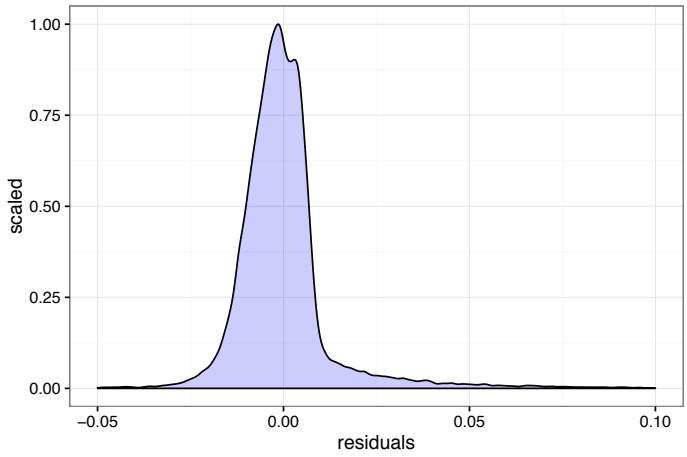

Figure 26: Density plot of linear model prediction residuals. A linear model was trained on LacZ- colonies to predict the 575nm absorbance in absence of CPR chromophore by using absorbance at 450nm and 650nm as input.

Using the above method, I was able to deconvolute the growth component from the CPRG turnover for each mutant across timepoints. CPRG turnover was calculated as the ratio of actual absorbance measurement at 575nm and the predicted growth-related measurement.

3.2.3.2 Separating mutants by CPRG turnover and growth fitness phenotypes

In order to examine whether CPRG turnover is a robust measurement of envelope integrity, I first calculated the reproducibility of CPR measurements across replicates for all timepoints. CPRG turnover varies dramatically in different timepoints, however CPR measurement across replicate colonies of the same mutant is quite reproducible when comparing data for the same timepoint (Pearson correlation r>0.79, see **Figure 27**).

At later timepoints there is an increased spread in CPR measurements, since the majority of the mutants are developing color, presumably resulting either from increased membrane permeability or cell death taking place during late stationary phase. Interestingly, this reveals several mutants that do not show any CPRG turnover at late timepoints, reproducibly behaving differently than the rest of the mutants. This underlines the potential of the assay to also identify genetic perturbations resulting in decreased lysis or membrane permeability (see also 3.2.3.4). A potential confounding factor of these measurements is the plasmid expression levels, this point is detailed in the Perspectives section (see 3.3.3).

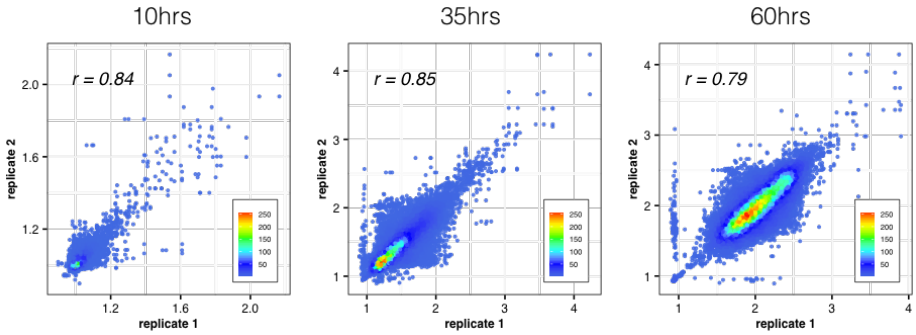

Figure 27: CPRG turnover measurement reproducibility across time; plot and Pearson correlation (r) were calculated in an all-against-all fashion measurement of 4 replicate colonies for each mutant in the library.

Simultaneously acquiring both growth and CPRG turnover measurements over time allows for side-by-side comparison of these two measurements. This in

turn can lead to interesting observations, such as the growth phase in which different mutants start to turn over CPRG. In other words, such timecourse observations can tell us which mutants either lyse or become permeable at the exponential phase, or at the early, or late stationary phase (see **Figure 28**B). As a general observation, most mutants seem to have an increased CPR turnover at the same time the growth rate slows down (see also 3.3.3).

An interesting observation is that the resolution of each of these phenotypes varies across timepoints. In fact, early timepoints have a greater resolution when it comes to growth fitness, measured as estimated absorbance at 575nm (see 3.2.3.1). This measurement saturates at later timepoints, since most mutants have reached stationary phase (see **Figure 28**A). However, these are the timepoints that have the highest dynamic range when it comes to measuring CPRG turnover.

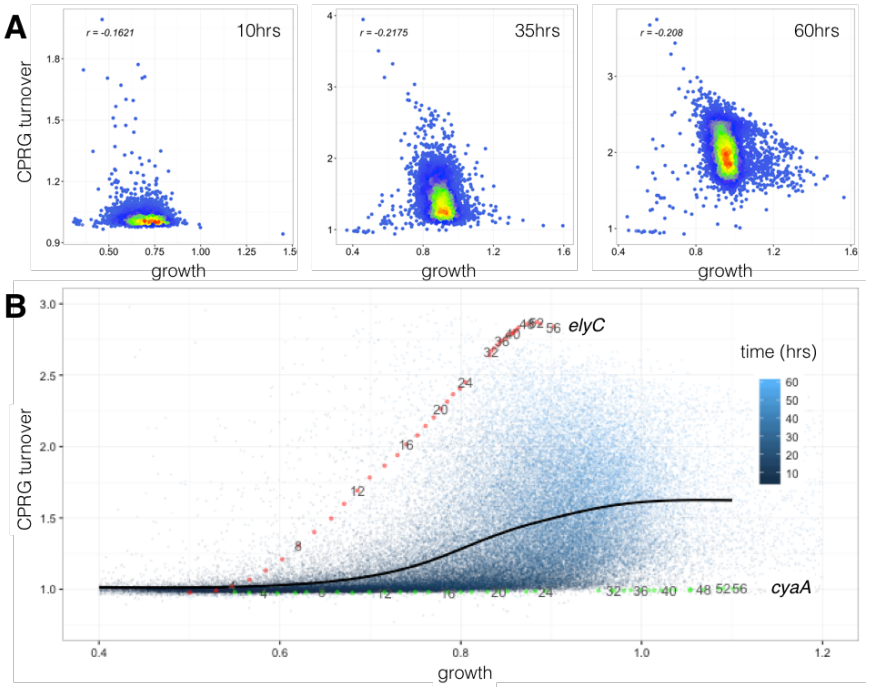

Figure 28: CPRG turnover compared to colony growth measurements across different timepoints. A. scatter plot in specific timepoints illustrates the progressive saturation in growth measurement and increase of CPRG turnover of most mutant colonies. B. CPRG and growth trajectories across timepoints of CPR+ *ΔelyC* mutant (red) and CPRG- mutant *ΔcyaA* (green); black line shows the result of a local non-parametric regression method (Cleveland, 1981)

Beyond monitoring specific timepoint snapshots, tracking bacterial colonies over time has the added advantage of accurate calculation of growth, as well as CPRG turnover kinetics. Such kinetic measurements can then be processed into calculating a growth rate, or a CPRG accumulation rate. Both of these tasks are

accomplished by means of robust fitting of a linear model on the largest linear part of a time series measurement.

While calculating such slopes is advantageous for acquiring a proxy of growth rates, using a similar approach for CPRG measurements would mean forfeiting late timepoint measurements, since these slopes are only linear for roughly the first 24 hours post-inoculation. Instead, for CPRG measurements I used the maximum CPRG component of the 575nm readout measured over 60 hours. This was done in order to avoid signal saturation (see example in **Figure 28**B), which I observed was the case for colonies that turn over CPRG very fast, presumably due to depletion of CPRG substrate (see 3.3.3).

3.2.3.3 Detection of mutants with elevated lysis or envelope defects

For each mutant in the library, I compared the CPRG turnover measurements of all 4 replicate colonies to the corresponding measurements of all mutant colonies. I performed this comparison by means of a two-sample t-test, which also takes into account the reproducibility of the measurements of each mutant. The t-test yields a probability that the mean of the two populations is equal, given the variance in these measurements. The populations tested are the measurements of all replicates of the mutant in question, and all replicate measurements of all mutants. Subsequently, I further processed these measurements to correct for multiple testing (Benjamini and Yekutieli, 2001), and acquired corrected p-values. Mutants with significant phenotypes are the ones that are found to have a corrected p-value lower than 0.05.

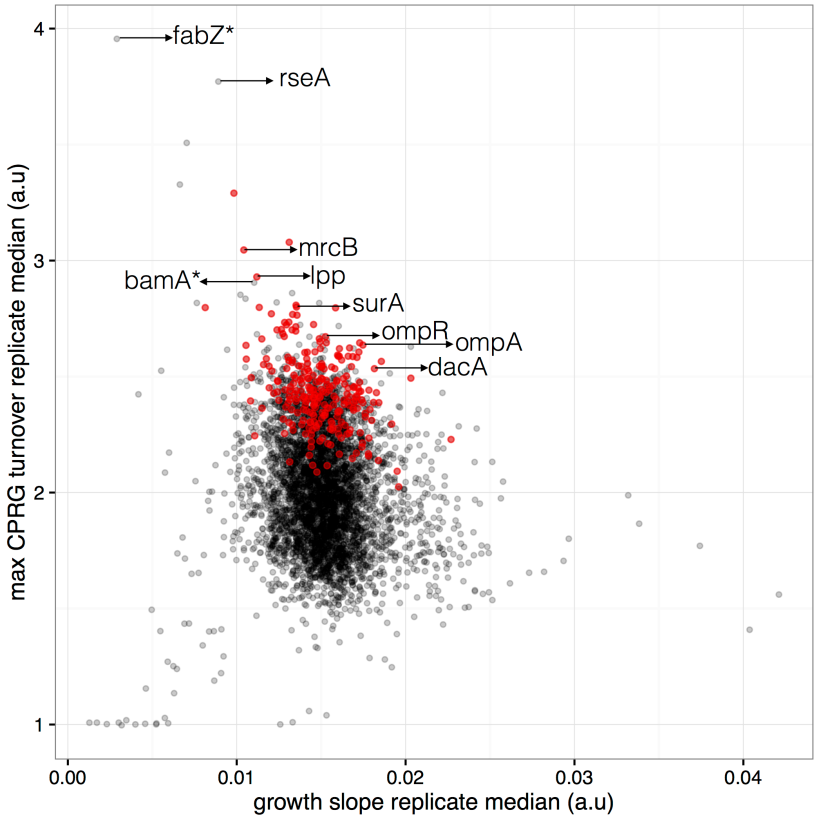

Figure 29: CPRG turnover versus growth rate (see 3.2.3.1). The *fabZ* SPA-tagged strain has a severe growth defect and a frequent lysis phenotype probably owing to defects in fatty acid synthesis (Heath and Rock, 1996). Red points indicate values corresponding to mutants that are found to have significantly higher CPRG phenotypes than the mutant population; these mutants are enriched for genes implicated in cell outer membrane assembly.

A GO enrichment analysis among the mutants with significantly increased CPRG turnover (see **Figure 29**), revealed a high enrichment in genes related to cell outer membrane, and outer membrane assembly (Benjamini-Hochberg corrected p-values: 0.009, and 0.035 respectively) (Benjamini and Hochberg, 1995).

Among the significant hits with higher CPRG turnover are three of the genes[1] of the β-barel assembly complex (BAM), which is required for assembly of β-barel outer membrane porins (OMP) (Hagan et al., 2011; Han et al., 2016). Also readily detected as significant hits are the well-studied OMP proteins OmpA, OmpG, and OmpF, the latter been shown to allow the passage of small solute molecules. SurA, also part of the significant hit list, is a chaperone protein implicated in the correct folding and transporting of OMPs, including OmpA and OmpF, from the inner to the outer membrane (Vertommen et al., 2009). SurA is known to interact with

[1] Gene *bamA* is essential; both reduced-activity versions of the gene in the

BAM, and a *ΔsurA* mutant was shown to exhibit elevated permeability presumably because of defects in OMP biogenesis (Behrens et al., 2001).

While defects in outer membrane (OM) are shown to increase membrane permeability or lysis frequency, other components of the cell envelope, namely the peptidoglycan layer (PG) and the inner membrane (IM) are equally as important to maintain cell integrity. Indeed Gram-negative bacteria need to coordinate this three-layered cell envelope during elongation and division. Lipoprotein Lpp, one of the most abundant proteins in *E. coli* is required for the stabilization of the cell envelope by physically tethering the OM to PG (Bernstein, 2011). Cells lacking Lpp have been shown to be more permeable to toxic compounds, while we find that a *Δlpp* mutant has significantly elevated CPRG turnover.

Defects in PG biosynthesis also have implications in cell permeability or lysis frequency. Specifically, we find that a *ΔmrcB* strain has elevated CPRG turnover, as also reported in (Paradis-Bleau et al., 2014). Gene *mrcB* codes for the PG biosynthesis protein PBP1B, a large macromolecular machinery spanning the periplasmic space, specialized in cell division in *E. coli*. Importantly, PBP1B was shown to interact with the Tol-Pal system to coordinate cell division and OM constriction (Gray et al., 2015). DacA, also known as PBP5, is a carboxypeptidase protein involved in PG processing and remodeling. While a *ΔdacA* strain has no growth phenotype (Matsuhashi et al., 1978), I find that cells lacking DacA either lyse more often or have a more permeable membrane.

3.2.3.4 Assay allows detection of mutant lysing less frequently or less permeable than the wild-type

Prolonged readout time allows the assay to uncover mutants which have reproducibly lower CPR turnover than the rest of the mutant population. This could either mean that such mutants either lyse less frequently, or that these mutants feature a more impermeable membrane than the wild-type, neither of which has been observed before.

Mutants that have significantly reduced CPR turnover (see Figure 30) are enriched for genes related to aerobic respiration and NADH dehydrogenase activity (*nuoBEF*), which is an integral part of respiration. Moreover, disruption of any of the genes encoding for the dipeptide ABC transporter Dpp (*dppABCDF*) results in drastically decreased levels of CPR turnover. Defects in the respiration chain will result in decreased levels of ATP, and thus decreased activity of ABC

transporters, such as Dpp. Taken together, these observations point to a possible active import of the CPRG compound by the Dpp transporter. This dipeptide transporter, as well as the oligopeptide ABC transporter Opp, were also recently implicated in the active import of the neglected antibiotics *kasugamycin* and *blasticidin S* (see (Shiver et al., 2016)). Further corroborating an active import mechanism of CPRG, most genes in the Opp operon also show decreased levels of CPR turnover.

There is the possibility that these promiscuous transporters also actively import IPTG, which is used for *lacZ* plasmid induction. However, IPTG is shown to enter the cell independent of an import system at higher concentrations (Marbach and Bettenbrock, 2012).

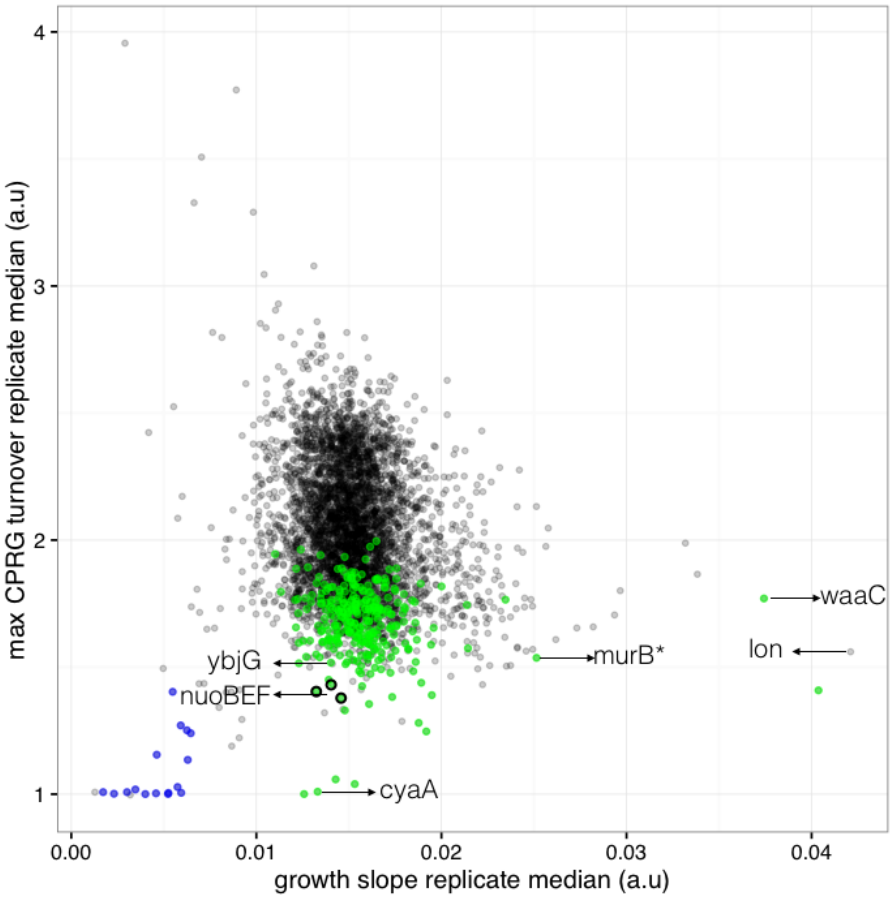

Figure 30: CPRG turnover versus growth rate (see 3.2.3.1). Green points indicate values corresponding to mutants that are found to have significantly lower CPR turnover than the mutant population; these mutants are enriched for genes implicated in aerobic respiration. Blue points correspond to mutants with growth defects and are excluded from the analysis.

An interesting result of the current screen is that genes involved in undecaprenyl phosphate (Und-P) recycling show a significantly decreased CPR turnover. Und-P is the universal lipid carrier used to export glycan components of carbohydrate polymers, such as PG, to the bacterial cell envelope. Und-P is derived from dephosphorylation of undecaprenyl pyrophosphate (Und-PP), a catalytic process required for Und-PP recycling after the transfer of the glycan component is complete (Tatar et al., 2007).

In *E. coli*, there are 4 proteins known to have Und-PP pyrophosphatase activity: *BacA*, *YbjG*, *LpxT*, and *PgpB*; *BacA* alone was shown to mediate 75% of this catalytic activity (Tatar et al., 2007). Deletion of genes coding for any of the first three enzymes unexpectedly results in significantly decreased CPR turnover. At the same time, a *ΔpgpB* strain showed increased CPR turnover, but PgpB is a multifunctional enzyme also involved in cardiolipin synthesis, one of the phospholipids composing bacterial membranes, which is shown to be regulated during osmoregulation (Romantsov et al., 2009).

Interestingly, deletion of a gene first identified by the preceding agar plate screen, *elyC,* also results in elevated CPR turnover and its role was pinpointed to be in Und-P utilization in PG biosynthesis (Paradis-Bleau et al., 2014). Further implicating specific PG biosynthetic steps, SPA-tagged versions (Zeghouf et al., 2004) of genes *murABCFGI* resulted in mutants that show a decreased CPR turnover (see Figure 32). These results and future experimental steps are further analyzed in section 3.3.2.

To rule out the possibility of spurious negative hits because of reduced enzyme levels, I and Matylda Zietek (Typas lab) tested several mutants for their β-galactosidase activity. While β-galactosidase activity after cell lysis varies among the mutants, results indicate that mutants that have significantly less CPR turnover do not have reduced β-galactosidase activity (see Figure 31).

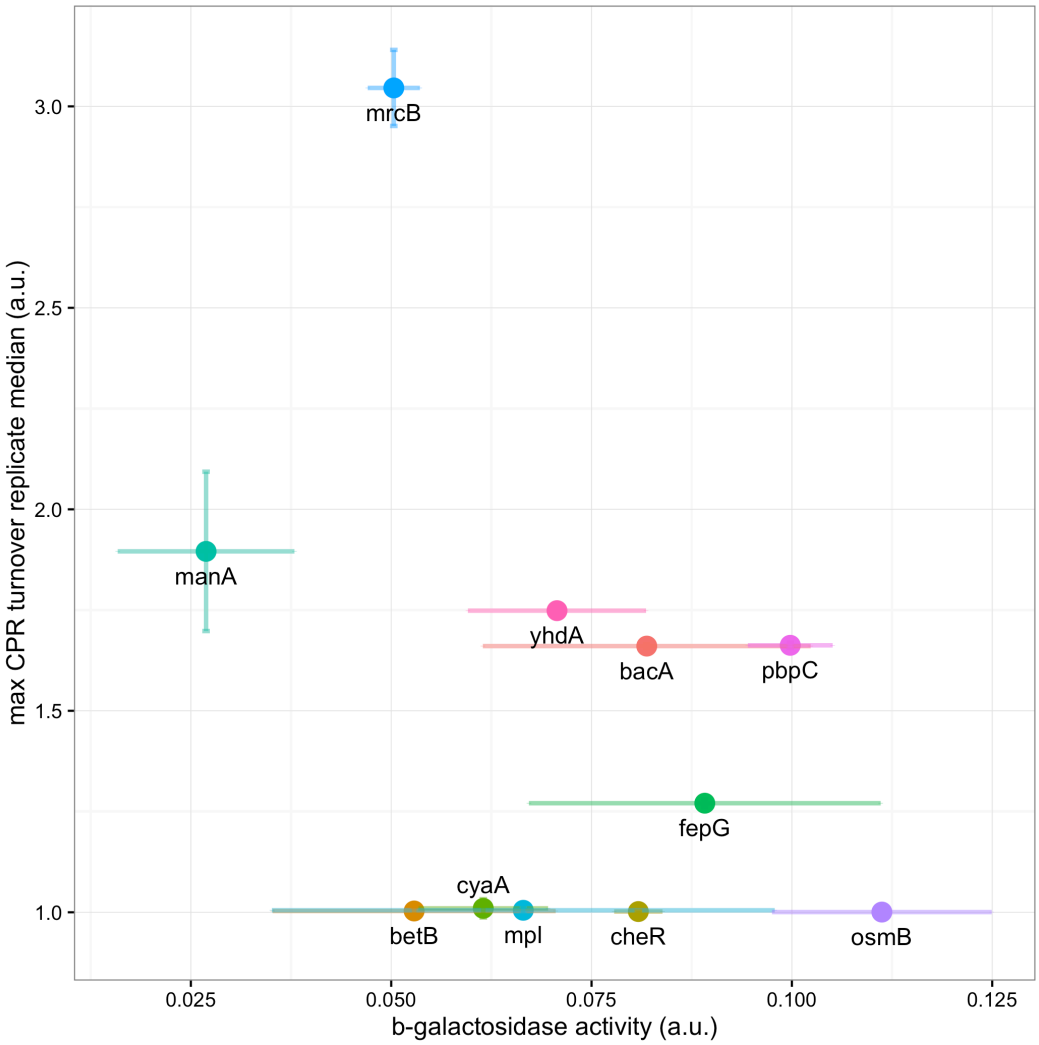

Figure 31: Testing for the β-galactosidase activity of CPR+ (top) and CPR- mutants (bottom). CPR- phenotype for these mutants is not attributed to decreased β-galactosidase activity. Also the CPR+ phenotype of the *ΔmrcB* strain is not due to increased β-galactosidase activity. Error bars denote one standard deviation away from the median of β-galactosidase activity (x-axis) and CPR turnover (y axis).

3.3 Perspectives

3.3.1 Method advantages

The described CPRG assay in well plates is a high throughput method to assess the degree of Gram-negative cell permeability or lysis frequency. An important advantage of this method is that it provides accurate quantitative measurements of both CPRG turnover, and colony growth. Moreover, most preparation and screening steps are automated, so the screen is easily scalable. Taken together, the current setup offers a quantitative measurement that is easy to

scale provide excellent grounds for accommodating a chemical genomics approach, whereby one can query the CPR turnover of entire bacterial libraries across chemicals.

Such CPR turnover measurements across genetic and chemical perturbations were also shown to provide information to that which is available from existing chemical genomics screens focusing on growth fitness. This allows us to make novel biological observations, concerning the frequency of lysis or cell envelope permeability across the exponential and stationary growth phases. Importantly, improvements in the assay have allowed us to mitigate technical issues in the preceding agar-plate assay and accurately measure CPR turnover for all mutants. These advance lead to the unprecedented observation of mutants that show significantly lower lysis or envelope permeability at stationary phase.

3.3.2 Disruption of specific envelope-related genes unexpectedly results in reduced CPR turnover

Reduced CPR turnover at stationary phase could mean that either the mutants have a reduced degree of cell lysis, or differences in the envelope allowing for reduced CPRG uptake. An unexpected result of the current screen is that mutations in specific genes involved in the biogenesis of cell envelope components results in reduced CPR turnover.

These mutants belong to the recycling of the Und-P lipid carrier, which is used to shuttle components of the (see Figure 32) PG sacculus and the enterobacterial common antigen (ECA) to the periplasmic side of the inner membrane. ECA is a surface polysaccharide produced by all enteric bacteria, however its function is still not known (Kuhn et al., 1988). Mutants of the ECA pathway were not found to have a strong CPR turnover phenotype in the current conditions. The PG and ECA biosynthetic pathways compete for use of the Und-P lipid carrier, and deletion of *wecA* has been shown to funnel all the Und-P towards PG biosynthesis, which alleviates the CPR+ phenotype of a *ΔelyC* strain (Paradis-Bleau et al., 2014).

Beyond the availability of Und-P due to decreased Und-PP recycling, mutations in genes involved in further steps of PG biosynthesis also result in decreased CPR turnover. SPA-tagged versions of proteins involved in the

assembly and the addition of peptides to the PG monomer (see Figure 32) all result in mutants with decreased CPR turnover levels.

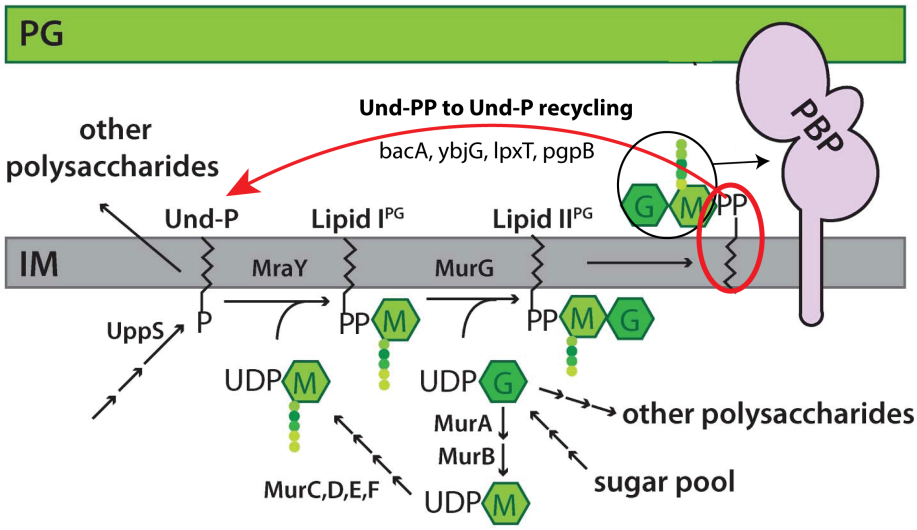

Figure 32: Peptidoglycan biosynthesis pathway. Mutants involved in lipid carrier Und-P recycling (red line), as well as PG components biosynthesis show a decreased CPR turnover. G, UDP-N-acetylglucosamine; M, UDP- N-acetylmuramic acid (see 1.1.1). Figure adapted from (Paradis-Bleau et al., 2014)

The trace of CPR- phenotypes continues to the final step of glycan strand polymerization and PG cross-linking. PBP1C is a large periplasmic protein that features both class-A PBP catalytic domains, transglycosylase (TG) and transpeptidase (TP), however its TP domain is shown to be inactive (Schiffer and Holtje, 1999). Lack of PBP1C does not have any effect on growth rate, or sensitivity on a range of antibiotics (Schiffer and Holtje, 1999), and is not expressed in exponential phase under lab conditions. However, I find that a *ΔpbpC* strain unexpectedly shows significantly decreased CPR turnover.

This observation comes in stark contrast to the strong CPR+ phenotype observed in a strain lacking another bifunctional PBP protein, PBP1B. Since PBP1B (*mrcB*) is involved in pre-septal PG formation, large part of the CPR+ phenotype of a *ΔmrcB* strain could be due to lysis during cell division. As expected, the lysis phenotype of a *ΔmrcB* strain is copied by a strain lacking LpoB, the lipoprotein activating PBP1B function (see Chapter 2). Further pointing towards a defect during lysis when PBP1B is missing or inactive, a strain missing the third bifunctional PBP of *E. coli*, PBP1A, has no observable CPR phenotype.

A noteworthy fact is that proteins PBP1B and PBP1C were shown in vitro to interact with each other, as well as with transpeptidase proteins PBP2 and PBP3 (Schiffer and Holtje, 1999; Vollmer et al., 1999). This suggests that the observed CPR phenotypes in deletion mutants could be either due to loss of function of the protein itself or malfunction of a potential protein complex encompassing these functions. In either case, measuring permeability phenotypes *en masse* sheds light on a yet-unobserved phenotyping aspect of these well-studied proteins.

These results indicate a specific part of the PG biosynthetic path that was up to now not implicated in cell lysis or envelope permeability. Currently, the mechanistic details of how these genetic perturbations can lead in reduced lysis or permeability are still elusive. A compelling hypothesis is to think such functions as hard-wired in the cell. In such a scenario, their normal activity into stationary phase could be part of programmed cell death for a fraction of the population, a well-known phenomenon in bacteria (Rice and Bayles, 2008; Sturges and Rettger, 1922).

One possible approach to aid our understanding of such a mechanism could be to combine genetic perturbations of this pathway. Meanwhile, a valuable control would be to test the expression levels of *uppS*, the gene responsible for generating Und-P de novo, in strains incapable of recycling Und-PP to Und-P (such as *ΔbacA*, *ΔybjG*, and *ΔlpxT*). In a related experiment, inducing the overexpression of *uppS* was shown to aleviate the CPR+ phenotype of the *ΔelyC* mutant, likely by increasing the available Und-P pool size (Paradis-Bleau et al., 2014).

3.3.3 Technical limitations

Both greatest strength and weakness of the current screening approach stems from the use of agar-filled microwell plates. The greatest technical challenge which was largely overcome to perform the current screen was the preparation of such agar well plates in a way that the agar level is uniform across all wells. Substantially different agar volume deposition in wells may affect the readout to a small degree, but more importantly could result in non-inoculated wells (see 6.2.3).

After optimizing the plate preparation protocol with Dr. Manuel Banzhaf (Typas Lab), the rate of non-inoculated wells is at 2% per plate. This allowed us to

inoculate all 4 replicates per mutant for most colonies, while 5% of mutants had 3 replicate colonies.

A technical factor possibly leading to decreased CPR turnover is reduced levels of the intracellular enzyme β-galactosidase. Decreased enzyme levels could be either due to decreased plasmid expression, or due to decreased plasmid copy number. Indeed a *ΔpcnB* mutant is known to have reduced number of plasmid copies per cell, and is readily detected as a significant CPR-negative hit in our screen.

To rule out the possibility of spurious negative hits because of reduced enzyme levels, I and Matylda Zietek (Typas lab) tested several mutants for their β-galactosidase activity (see results in 3.2.3.4, and Figure 31). As a future step, I plan to test for β-galactosidase enzyme activity in a high-throughput manner across all mutants in the assayed bacterial library. Such data will then help correct CPR turnover phenotypes in each mutant for their β-galactosidase activity.

Another technical limitation of the present approach may arise because of differences in the *lacZ* plasmid expression across different growth phases. In fact, we found that the current plasmid used for the screen (pCB112, see 6.2.1) has higher expression levels at stationary phase. Results are still comparable, since at late timepoints all mutant colonies should reach stationary phase, save for mutants with severe growth defects which are excluded from the analysis. In any case, this issue could be causing loss of information in the exponential growth phase. I thus plan to perform more testing steps using different plasmids to mitigate this issue. Mutant colonies that show a significant growth defect

Also subject to further testing and optimization is the amount of substrate in each agar-filled well. In mutants that have a very high-level of CPR turnover (such as *ΔelyC* or *ΔmrcB*) I observed signal saturation after several hours of incubation. To understand whether this is due to a biological phenomenon, or attributed to substrate depletion, I plan to perform further testing using varying amounts of CPRG and β-galactosidase levels.

3.3.4 Comparison of two screening approaches

The described CPRG assay in well plates was conducted at room temperature in LB containing 1% NaCl. This was done to mirror one of the conditions used in

the imaging-based agar plate and thereby allowing me to compare the current results with results acquired using the classic agar plates.

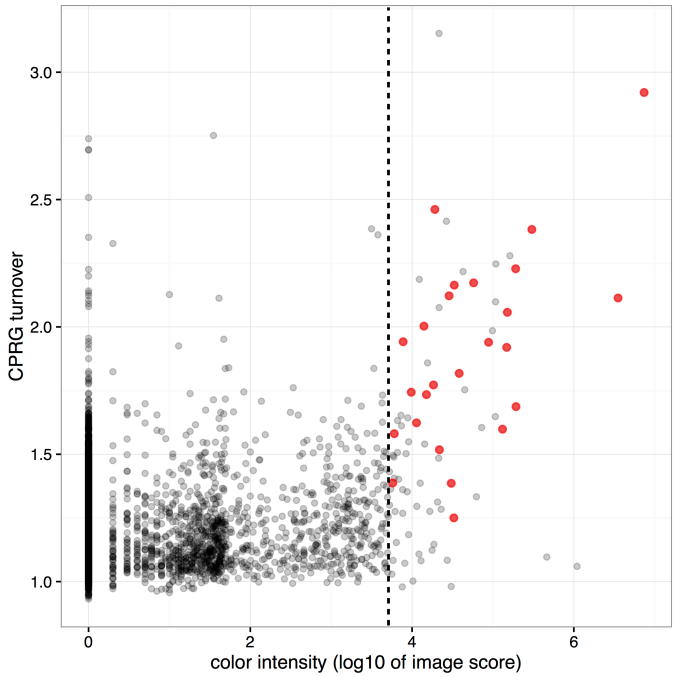

Figure 33: Comparison of CPRG turnover in agar-well plates values and agar plates from the preceding screen at the 24 hours post-inoculation timepoint. Red points indicate mutants that are CPR+ hits from both screening methods. Most mutants over the threshold of 3.7 in agar plates (dotted line, log10 scale) already show a high CPRG turnover in agar-well plates at 24 hours.

As expected the prolonged measurement time only possible in agar well plates enables us to uncover hits, of which were not be present on agar plates. In addition, the limited development time of agar plates makes the readout susceptible to variations stemming from differences in inoculum. Such differences could account for the high number of false-negatives in the agar plate assay. To avoid spurious hits, a strict cutoff was set at 3.7 (log scale, see **Figure 33**) in the published data. Most hits over this threshold have very high CPR turnover levels measured in the current well-plate setup.

3.3.5 Outlook

3.3.5.1 Envelope permeability chemical genomics

After solving remaining technical challenges (see 3.3.3), the current approach for measuring membrane permeability or lysis rate can be easily scaled to allow testing across chemical perturbations. Importantly, the provided readout is both quantitative and reproducible across several replicates. These qualities make the current high-throughput screening approach amenable to be applied across chemical conditions. Indeed, one can select compounds to serve as perturbations on different components of the bacterial cell envelope. Such an approach combines targeted chemical perturbations with the preexisting genetic perturbation of gene deletion strains.

Importantly, the data acquired in the imaging-based agar plate screen clearly demonstrate that key mechanisms involved in cell envelope integrity may change according to environmental conditions. This result follows the small overlap of hit genes across the different temperature and media osmolarity conditions. Thus, I expect a chemical genomics screening approach to uncover a range of mechanisms involved in the cell reaction to chemical insults against its envelope.

Moreover, such data allow the creation of an envelope integrity profile for each mutant across conditions. As in other chemical genomics approaches (Nichols et al., 2011), a pairwise similarity of mutant profiles can then easily be calculated. This can in turn lead to hypothesis generation in a guilt-by-association manner. For example mutants that behave similar across conditions are possibly involved in the same process or pathway.

Last but not least, such envelope integrity chemical genomics data can be readily used to complement existing high-throughput chemical genomics data on growth fitness (Nichols et al., 2011). As an example, drug sensitivity of a particular mutant can then be further dissected into growth arrest or lysis. By combining such data across conditions will thus lend a new level of detail in observing mutant condition-specific phenotypes, possibly leading to better understanding of growth inhibition and drug mode-of-action.

3.3.5.2 Envelope screening across species

An envelope integrity chemical genomics screen can provide a wealth of data that will increase our understanding of the molecular mechanisms underpinning the synthesis of the Gram-negative cell envelope. At the same time, such an approach can aid uncover potential drug targets, to help overcome the molecular barrier posed by the cell envelope, a limiting factor in drug treatment options during Gram-negative bacterial infections.

Such a screening approach described here can readily be adapted to screening mutant sets across variable Gram-negative species. For example, the model pathogen bacterium *Salmonella enterica* is amenable to the same genetic tools as *E. coli* allowing for mass incorporation of a LacZ-encoding plasmid to facilitate high-throughput envelope integrity screening. Moreover, the model pathogen *Pseudomonas aeruginosa* exhibits dedicated nutrient uptake porins, as well as more efflux systems than *E. coli*, which make it easily adaptable for drug resistance. Additionally, there is a readily available transposon mutant library for *P. aeruginosa* strain PAO1, which already features the *lacZ* gene (Jacobs et al., 2003).

Envelope integrity data across Gram-negative species will eventually facilitate cross-species comparison. Comparing the envelope integrity profile of well-studied genes to unknown function genes in other species can help elucidate their molecular role. Since this assay is very sensitive in detecting lysis, a potential interesting application of this assay would also be as a readout in cross-species interactions, as well as phage, or toxin attacks to the bacterium.

3.4 Contributions

Experiments and analysis presented in this chapter is divided into two projects. Results of the preceding image-based envelope integrity screen have already been published in (Paradis-Bleau et al., 2014). My role in that project was to develop the image analysis software that provides the CPRG readout on agar plates. I also performed the downstream data analysis. Following that I improved this assay by the use of well plates. Optimization of the agar well plate preparation and inoculation was performed by Dr. Manuel Banzhaf, Anja Telzerow, Lucia Herrera (Typas lab), and myself. Together with Dr. Manuel Banzhaf we performed the well-plate preparation and experiments detailed in the latter part of this chapter. The β-galactosidase activity assay was performed by Matylda Zietek

(Typas lab). Finally, I performed the absorbance modeling-based readout optimization, and developed the software tools to analyze the data.

4 Coevolution of domains in modular proteins

4.1 Background and significance

The peptidoglycan mesh is a major ubiquitous component in bacterial cell walls. This rigid mesh helps bacterial cells maintain their shape, and the ability to withstand osmotic stress, given their own turgor pressure. Key indispensible enzymes required for cell wall biosynthesis are proteins called Penicillin-Binding-Proteins (PBPs). Two of the *E. coli* PBPs (PBP1A and PBP1B) harbor domains that serve as docking regions for cognate lipoproteins that activate the PBP enzymatic function upon binding to their respective docking domains (Egan et al., 2014; Paradis-Bleau et al., 2010; Typas et al., 2010). Following the observation that these domains, and their cognate lipoproteins co-occur across species (Typas et al., 2010), I set forth to explore the domain content plasticity in core process proteins, and especially those involved in cell wall growth, with the aim of uncovering the interacting partners of these domains.

In this chapter, I make extensive use of the terms *domain,* and *conserved region*. Conserved regions are protein sequence regions that are found to retain high sequence similarity across bacterial strains or species. Protein domains are a subset of conserved regions that are annotated in specialized databases (Finn et al., 2014). Protein domains are often shown to evolve, and function separately from the remaining protein (Apic et al., 2001). While new domains are continuously added to databases, in this study I will refer to domains as those conserved regions that have been annotated as domains at the time of writing.

4.1.1 Bacterial cell wall biosynthesis is a modular process

The peptidoglycan layer is synthesized by an enzyme class termed Penicillin-Binding-Proteins (PBPs). Bifunctional PBPs catalyze the elongation of glycan strands, and perform the glycan strand crosslinking to create the peptidoglycan (PG) mesh. Glycan strand elongation is performed by the glycosyltransferase (GT) protein domain, while cross-linking strands is done by the transpeptidase (TP)

domain. PBPs can be either mono- or bi-functional, meaning they harbor one or both of these catalytic domains.

While GT and TP functions are important for PG biosynthesis, a range of heterogeneous molecular functions are required to process the PG layer in order for the cell to remain functional. For instance controlled degradation of PG cell wall is required for inserting new nascent glycan strands, as well as insertion of macromolecular machineries that span the periplasm. To this end, a host of enzymes are responsible for processing the glycan strands and the peptides used for strand cross-linking. Peptide crosslinks are cleaved by endopeptidases, while the peptides are trimmed by carboxypeptidases or removed from the glycan strands by amidases. Finally, lytic transglycosylases process and terminate the glycan strands (Typas et al., 2012).

Interestingly, each such function typically features a high level of redundancy, with several proteins undertaking the same or similar tasks. For example, *E. coli* features 6 lytic transglycosylase and 4 amidase enzymes. This redundancy is thought to have a double role, contributing both in robustness of the process as well as specialization, so it can integrate input from different processes. For instance, deleting any single gene of most of these enzymes in *E. coli* has little or no effect to cell shape or growth rate. On the other hand, there is accumulating evidence supporting the view that these different enzymes are specialized for remodeling the cell wall and adjusting its elasticity and other properties at different growth phases or cell division stages (Yahashiri et al., 2015).

This functional redundancy and yet specialization of related cell wall enzymes largely reflects the modularity in their domain content. Redundant or functionally related enzymes share the same catalytic domains, yet the rest of their sequences bear little to no similarity. Moreover, this surrounding sequence is frequently conserved among homologues of the same protein in closely related species, but lost quickly in evolution.

Recent evidence suggests this plasticity in protein non-catalytic regions is closely intertwined with regulation of their enzymatic function. For instance, proteins PBP1A and PBP1B of *E. coli* are mutually redundant, yet they are specialized in peptidoglycan biosynthesis during cell elongation and cell division respectively. For these two enzymes it was recently shown that domains and regions outside the catalytic domains play a pivotal role in the regulation of their function (Egan et al., 2014; Markovski et al., 2016; Paradis-Bleau et al., 2010; Typas et al., 2010).

4.1.2 Coevolution of protein domain and interacting partner

Recent work has shown the importance of a pair of lipoproteins, LpoA and LpoB in the activation of their cognate Penicillin Binding Proteins, PBP1A and PBP1B respectively (Egan et al., 2014; Markovski et al., 2016; Paradis-Bleau et al., 2010; Typas et al., 2010). In 2010, Typas et al. and Paradis-Bleau et al. independently found that lipoproteins LpoA and LpoB physically interact with their cognate PBPs and activate their function. Later, the LpoB cofactor protein was shown to span the periplasm and bind to a specific conserved region within PBP1B (Egan et al., 2014). More recent work revealed amino-acid substitution variants of the *E. coli* PBP1B that bypass the need for LpoB activation. These findings support a model where LpoB upon binding to the UB2H domain induces a conformational change to PBP1B, activating its Glycosyltransferase (GT) domain, which in turn also increases its TP activity (Markovski et al., 2016).

Meanwhile, the LpoA cofactor protein was shown to directly bind to the ODD region of PBP1A in *E. coli*, both in vitro and in vivo (Typas et al., 2010). Conversely with the activation of the GT activity of PBP1B, LpoA was shown to activate its cognate PBP by a different mechanism. LpoA primarily activates the TP activity of PBP1A, which then also results in increased GT activity (Egan et al., 2015; Lupoli et al., 2014).

Importantly, these cofactor proteins were shown to follow the phylogenetic distribution of the binding regions in their corresponding PBPs. Although bifunctional PBPs are widespread among bacteria, only the PBPs in species within Gamma-proteobacteria have a conserved ODD region. The exact same set of species was found to feature the LpoA protein in their genome. Interestingly, LpoB exhibits a similar behavior, since it follows the same phylogenetic distribution as the UB2H domain on PBP1B. In the case of the UB2H-LpoB pair, both domain and cognate protein were confined within Enterobacteria (Typas et al., 2010).

Taken together these results indicate that cofactor proteins such as LpoA and LpoB co-evolved with their cognate binding regions in PBPs. Interestingly, the LpoB-mediated regulation of PBP1B is more evolutionary confined than the LpoA-ODD regulation of PBP1A. This suggests that this coevolution of domain and interacting partner has occurred at least twice within Gamma-proteobacteria.

Furthermore, this implies that niche adaptation may have contributed to these differences in peptidoglycan synthase regulation across phyla.

Highlighting the importance of this cofactor-binding region coevolution, a recent study showed that another cofactor protein, also physically interacts with the UB2H region of the *E. coli* PBP1B (Gray et al., 2015). This cofactor protein named CpoB, corresponds to the last gene in the Tol-Pal operon. In the same study it was demonstrated that CpoB also interacts with the Tol machinery to coordinate peptidoglycan synthesis at the division site. These results indicate that such conserved regions play an important role in the spatio-temporal regulation of cell wall biosynthesis machineries.

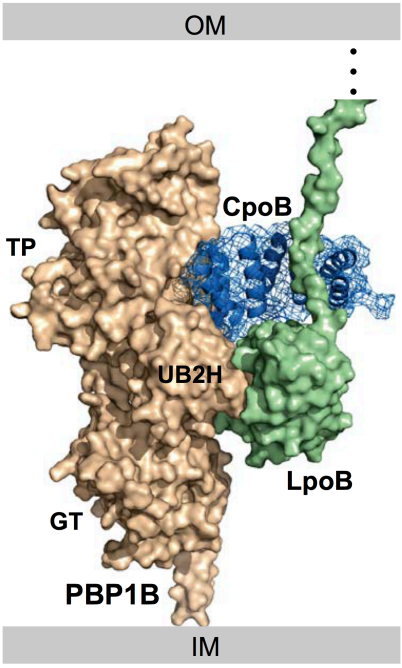

Figure 34: Model for physical interaction of the PBP1B-LpoB-CpoB complex. Showing the lowest energy conformation calculated by docking algorithm HADDOCK (de Vries et al., 2010) after integrating experimental data (Gray et al., 2015). LpoB binds the UB2H domain, while CpoB binds in a cleft between UB2H and the TP domain in PBP1B. Figure copyright (Gray et al., 2015), licensed under CC-BY.

4.1.3 Domain content plasticity across evolution

Protein domains are conserved parts of a protein sequence that are often shown to evolve, function, and exist independently of the rest of the protein. Moreover, domains are viewed as building blocks that may be recombined in different arrangements to create proteins with different functions (Apic et al.,

2001). Proteins often consist of several domains, while the same domain may appear in a variety of proteins.

While domain content varies among proteins of overlapping catalytic functions, domain content also varies across evolution. In homologues of the same protein across species, we often observe the same catalytic domain content being widely conserved, while the non-catalytic domain content changes quite rapidly. Such examples are the bifunctional PBPs, which can be found across the bacterial kingdom. The catalytic domain of bifunctional PBPs stays the same, but their non-catalytic domain content varies significantly across evolution (see Figure 35).

In more detail, non-catalytic domains and certain regions framing catalytic domains tend to be confined within contiguous parts of the bacterial phylogenetic tree. Conservation of these non-catalytic domains and regions also varies widely: from confined between only a few related species to spanning multiple phylogenetic classes.

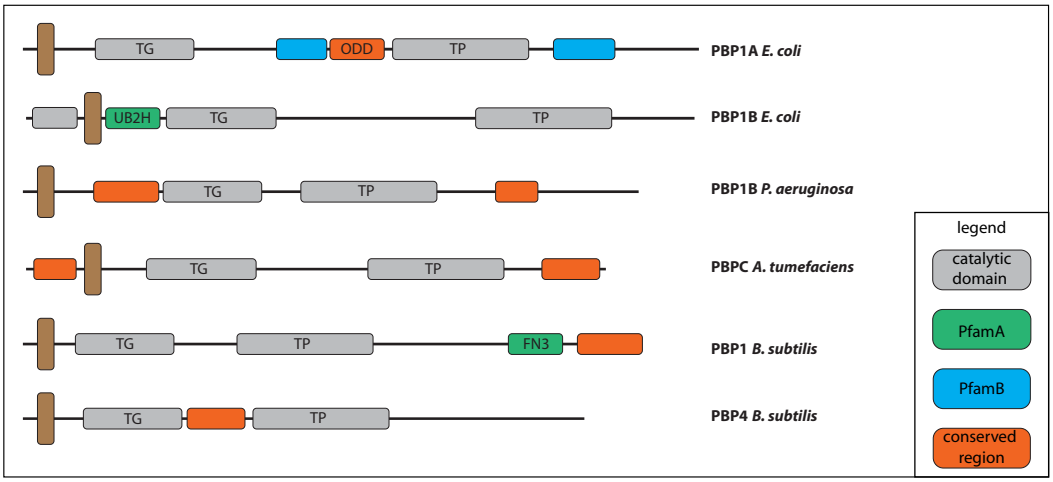

Figure 35: Domain annotations in bifunctional PBP proteins across species illustrates their non-catalytic domain content plasticity both within species and across evolution

4.1.4 Using coevolution to detect domain-protein interactions

Lipoprotein cofactors in *E. coli* were shown to regulate the function of their cognate PBPs by binding to conserved regions within those PBPs. Both cofactor protein and region feature the same phylogenetic distribution suggesting that regulators coevolved with their binding regions. Moreover, this was shown to have

occurred at least twice in evolution with different levels of conservation. This last observation implies that regulator coevolution to binding region in a modular protein may be a more extensive mechanism to spatio-temporaly regulate the function of core process enzymes, as part of niche adaptation.

Thus exploring the coevolution signal of domains in core process enzymes may hold the potential of uncovering possible regulator cofactors. The periplasmic space offers a particularly good starting point for this study, since it is devoid of ATP. Thus, regulation of periplasmic processes often occurs via post-tranlational procedures, including direct physical interactions and protein complexes.

Exploring protein-protein coevolution to uncover possible interaction partners is a decade-old field, with many proposed computational tools. Molecular coevolution methods are generally subdivided in whole-protein coevolution (Pazos and Valencia, 2001; Szklarczyk et al., 2014; Tillier and Charlebois, 2009), and residue-level coevolution methods (Jones et al., 2012; Reynolds et al., 2011; Yang et al., 2011). However different in application, the whole spectrum of molecular coevolution methods is based on the observation that interdependent molecules or residues will tend to change in a correlated way across evolution.

Residue-level coevolution methods typically aim to discover functional associations between residues in the same protein molecule. These methods invariably use the multiple sequence alignment of protein homologues across species to search for correlated mutations. Such correlated mutations have been shown to occur between residues of physical proximity or those acting together in catalytically sites. Recently, residue-level methods have been extended to encompass groups of coevolving residues, which were shown to mediate allosteric regulation between distant parts of a protein (Reynolds et al., 2011).

Protein-level coevolution methods are split between methods that use the similarity of phylogenetic trees, and those that use the similarity of protein phylogenetic distribution (or phylogenetic profile). The former methods compare the distances in the phylogenetic trees of two proteins, following the observation that phylogenetic trees of functionally related proteins tend to have a high degree of similarity. Instead, phylogenetic distribution-based methods rely on the observation that functionally related proteins tend to follow the same pattern of presence or absence across species.

Domain-level coevolution on the other hand is the obvious middle-ground between those divisions of molecular coevolution. A series of studies (Luo et al.,

2011; Pagel et al., 2006; Pagel et al., 2008) addressed domain-domain coevolution by comparing the phylogenetic profiles of all the domains in the Pfam database and integrating those to existing experimental information available in the iPfam database (Finn et al., 2014). Although of great importance, this work was not followed up on and has not been updated since September 2010, thus only uses a fraction of the sequences and domain annotations currently available, which are covered within this work.

Moreover, several screening approaches have focused in elucidating the interacting domains of proteins. This is typically performed by using yeast-2-hybrid (Y2H) to observe which fragments of two known interacting proteins also interact (Ryan et al., 2013). However, no new work has been published on computational methods to predict domain-domain interactions, excluding work that aggregates existing domain-domain coevolution information (Kim and Mylonakis, 2011; Memišević et al., 2013). One probable reason for this are the technical challenges in adapting existing protein-level coevolution techniques to protein domains. In fact, many of these methods are very resource-intensive both for computational time, and physical memory required.

Finally, to the best of my knowledge, there is no published work describing methods to explore conserved region or domain coevolution with a protein. The modular proteins I focus on this study hold multiple conserved regions and domains. Thus detecting coevolution signals among domains would require scaling up the required computational time by two orders of magnitude (see section 4.3.1.4), making it prohibitive for most computational infrastructures. In this study I confine the search for possible protein cofactors to conserved regions and domains in proteins involved in cell wall biosynthesis and remodeling. On the positive side, limiting the number of query regions allows me to increase the search space for proteins of each such region.

4.2 Results

First I collected a list of proteins in *E. coli* that are known to be involved in cell wall biosynthesis and remodeling (see Table 1). I then analyzed their domain content, to assemble a list of catalytic domains that correspond to a comprehensive set of functions needed for correct cell wall. Next, using each one

of these domains, I searched across a set of 20 phyloegenetically divese species for proteins that hold such enzymatic domains. These species were selected on account of their utility in molecular biology and biotechnology, and because they collectively span phylogenetically the entire bacterial kingdom (see Table 2).

Table 1: list of cell wall-related proteins in *E. coli* and their enzymatic functions

function	proteins
1,6-anhydro-N-acetylmuramyl-L-alanine amidase	AmpD
Beta-lactamase	AmpC
Bifunctional Penicillin-binding protein	MrcA, MrcB, PbpC
Cell division protein FtsA	FtsA, FtsZ
Cell shape-determining protein	MreB, MreC, MreD, MrdB, RodZ
D-alanyl-D-alanine carboxypeptidase	DacA, DacB, DacC, DacD, PbpG
DD-carboxypeptidase	YfeW
glycosyltransferase (GT)	MurG
Inhibitor of lytic transglycosylases	Ivy
Lipid II flippase	MurJ
lytic transglycosylase	MltA, MltB, MltC, MltD, MltE, MltF, Slt
Monofunctional biosynthetic peptidoglycan transglycosylase	MtgA
Monofunctional Penicillin-binding protein (TP)	FtsI, MrdA
Murein DD-endopeptidase	MepH, MepM, MepS, DacA, PbpG
Murein hydrolase activator	EnvC, NlpD
N-acetylmuramoyl-L-alanine amidase	AmiA, AmiB, AmiC, AmiD
Penicillin-binding protein 2	MrdA
Penicillin-binding protein activator	LpoA, LpoB
Penicillin-insensitive murein endopeptidase	MepA
Probable endopeptidase	NlpC, YafL
Probable L,D-transpeptidase	YbiS, YcbB, YcfS, ErfK, YnhG, YafK
SEDS (shape, elongation, division, sporulation)	FtsW, RodA
unknown	NlpI, YgeR

Table 2: List of species selected to span the bacterial kingdom. The domain content of all proteins containing cell wall-related enzymatic domains in these species was analyzed to search for their potential co-evolving partner proteins.

strain name	class	order
Escherichia coli K-12	Gamma-proteobacteria	Enterobacteriales
Haemophilus influenzae Rd KW20	Gamma-proteobacteria	Pasteurellales
Vibrio cholerae O1 biovar El Tor str. N16961	Gamma-proteobacteria	Vibrionales
Shewanella oneidensis MR-1	Gamma-proteobacteria	Alteromonadales
Pseudomonas aeruginosa PAO1	Gamma-proteobacteria	Pseudomonadales
Neisseria gonorrhoeae FA 1090	Beta-proteobacteria	Neisseriales
Caulobacter crescentus CB15	Alpha-proteobacteria	Caulobacterales
Agrobacterium fabrum str. C58	Alpha-proteobacteria	Rhizobiales
Myxococcus xanthus DK 1622	Delta-proteobacteria	Myxococcales
Desulfovibrio vulgaris str. 'Miyazaki F'	Delta-proteobacteria	Desulfovibrionales
Streptococcus pneumoniae R6	Firmicutes	Lactobacillales
Lactobacillus acidophilus NCFM	Firmicutes	Lactobacillales
Staphylococcus aureus subsp. aureus NCTC 8325	Firmicutes	Bacillales
Bacillus subtilis subsp. subtilis str. 168	Firmicutes	Bacillales
Clostridium perfringens str. 13	Firmicutes	Clostridiales
Streptomyces coelicolor A3(2)	Actinobacteria	Streptomycetales
Mycobacterium tuberculosis H37Rv	Actinobacteria	Corynebacteriales
Synechocystis sp. PCC 6803 substr. Kazusa	Cyanobacteria	Chroococcales
Bacteroides thetaiotaomicron VPI-5482	Bacteroidia	Bacteroidales
Helicobacter pylori 26695	Epsilon-proteobacteria	Campylobacterales

For each one of the proteins in these 20 species that were found to hold a catalytic domain of interest, I then analyzed its domain content to acquire protein regions and domains that are not predicted to be catalytically active.

Then, following the pipeline I developed (steps 1-3, see section 4.2.1) I was able to retrieve the phylogenetic distribution, and domain context of more than 150 PfamA, PfamB domains, and conserved inter-domain regions. I discarded poorly conserved domains and regions, as well as those that are promiscuous for the catalytic domain they are found next to. This step helped refine the focus to 97 PfamA, PfamB domains and regions[2]. These protein domains and regions were further processed in the subsequent steps of the pipeline (steps 4-6, see section 4.2.1), and their results are detailed below. Since finding and verifying region conservation is still dependent on manual steps, I have systematically explored only the conserved regions that can be found in *E. coli*. Provided this approach is

[2] 19 PfamA domains, 29 PfamB domains, and 49 conserved regions. Conserved regions are only those that can be found in *E. coli* proteins.

fruitful for these regions, there are several more regions conserved in other parts of the phylogenetic tree that are left to explore.

In the present document, I will use the ODD of PBP1A together with its cognate lipoprotein LpoA as a positive control. This positive control pair serves as an example; results of the pipeline when invoked with other domains in the positive controls are similar but cannot show all of them due to space limitations.

4.2.1 Pipeline to detect domain-protein interaction partners

Following the observation that domains in modular PBPs co-occur across species with their cognate lipoproteins, I explored the power a domain-protein co-occurrence approach would have with respect to the positive control domain-protein pairs. To this end, I explored an extensive set of domains, including the three domains in the positive control set, using the following approach:

1. I first compiled a list of *E. coli* proteins that are central to the bacterial core process of cell wall biosynthesis and remodeling.
2. For each of the proteins in this list, I acquired all their homologues across 20 different species that span the whole bacterial kingdom.
3. I then acquired a comprehensive set of domains and inter-domain regions[3] that are found in these proteins and their homologues, either annotated as catalytically active or not.
4. I explored the phylogenetic distribution of each of the non-catalytically annotated domains and regions, only when they are found on the same protein featuring the catalytic domain content of the original protein.
5. I compared the phylogenetic distribution for each one of these domains and regions to the phylogenetic distribution of all clusters of orthologous genes (COG) in the EggNOG database.
6. Finally, I compared the phylogenetic trees of these domains to those of every COG in the EggNOG database

[3] There are numerous such regions in cell-wall related proteins across the species examined in this work. However, quality control for these regions involves manual steps, so the conserved regions explored in this work are limited to the ones that can also be found in proteins *E. coli*.

As an intermediate result, at the second-to-last step I can already rank proteins (corresponding to COGs) according to the degree of phylogenetic distribution similarity to the domain of interest.

4.2.2 Co-occurrence measures predict known co-evolving domain-protein pairs

Following the paradigm of the positive control domain-protein pairs, I first explored the potential of an approach that would detect protein-domain coevolution by means of their co-occurrence across species. Indeed comparing the phylogenetic distribution of any positive control yields high degree of similarity. For example, by comparing the phylogenetic distribution of the PBP1A ODD domain to its cognate lipoprotein LpoA, the overlap in co-occurring species yields an F-measure of 0.8755 and Mutual Information of 0.882. In order to assess the predictive power of this co-occurrence approach, I acquired co-occurrence values for all clusters of orthologous genes in the EggNOG database (Huerta-Cepas et al., 2016).

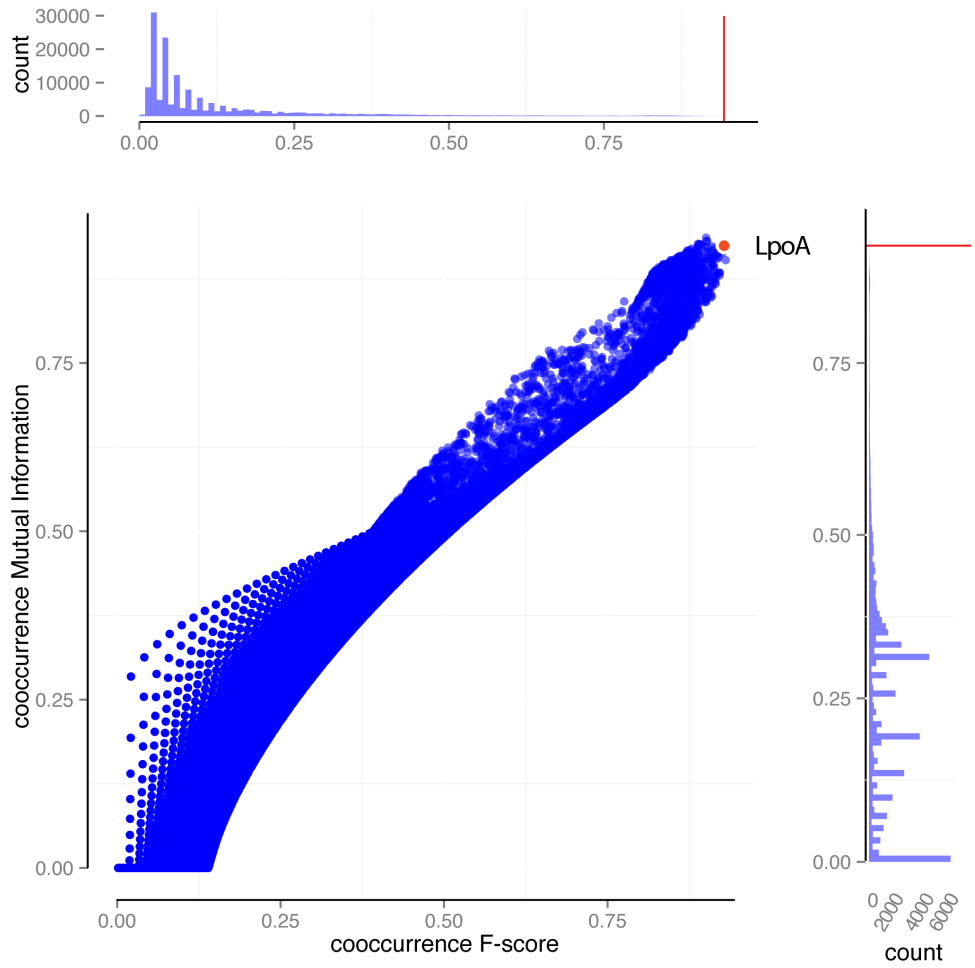

Figure 36: comparison between F2-measure and Mutual Information (MI), ODD domain tested phylogenetic distribution tested against all EggNOG gene clusters, indicated gene cluster corresponds to LpoA, the lipoprotein cognate to the ODD domain.

Typically the gene clusters corresponding to positive control proteins rank within the top 0.2% when comparing their phylogenetic distribution to the ones of their cognate domain. Whilst these positive control gene clusters are typically not the first hits, one needs to take into account that I tested all gene clusters in the eggNOG database, exceeding 104.000 gene clusters. In fact, by modeling the F-measure distribution of all gene clusters by a beta distribution, the LpoA gene cluster has a P-value of 2.4×10^{-6} (see Figure 37).

Finally, as shown in Figure 36, co-occurrence measures are redundant to an extent. However, this redundancy is not as pronounced in very high values (over 0.85), where our positive control gene clusters usually rank. As demonstrated in section 4.2.2.2, gene clusters that are highly-ranked by co-occurrence typically belong in the same process as the query domain. Thus, using the result of both co-

occurrence measures helps in better ranking of known positive control gene clusters compared to gene clusters corresponding to proteins that are simply part of the same process.

4.2.2.1 Filtering by localization improves method specificity

This co-occurrence method achieved a very high predictive potential for all 3 known interacting domain-protein pairs. I wanted to explore whether applying filters based on existing knowledge on the hits would improve the ranking of hits. For example, positive control proteins corresponding to domains are all periplasmic lipoproteins. In fact, most of the modular proteins I explored in the scope of this study are localized in the periplasm, or anchored to the inner membrane but with domains in the periplasm. Thus, such filtering for proteins that co-localize is expected to improve the specificity of my method without sacrificing sensitivity. Indeed, filtering results for proteins that localize in the periplasm (see 6.3.8) yields far better results for the positive controls:

- ODD-invoked pipeline ranks LpoA as 1st hit,
- UB2H invoked pipeline ranks LpoB as 8th hit
- BiPBP_C invoked pipeline ranks YfhM as 1st hit.

4.2.2.2 Co-occurrence measures enrich for proteins in the same process

Exploring the top hits of the list, I observed that a lot of the genes that are known to be part of the same process as the domain I use as the sole query input. In order to query whether this is a general trend, I performed a Gene Ontology (GO) enrichment analysis to the top 500 gene clusters, ranked by co-occurrence F-measure. These 500 gene clusters correspond to the top 0.5% of all gene clusters. By far the most enriched process in the top 0.5% of all gene clusters is peptidoglycan biosynthetic process, featuring a Benjamini-Hochberg-corrected P-value of 2.1×10^{-6}.

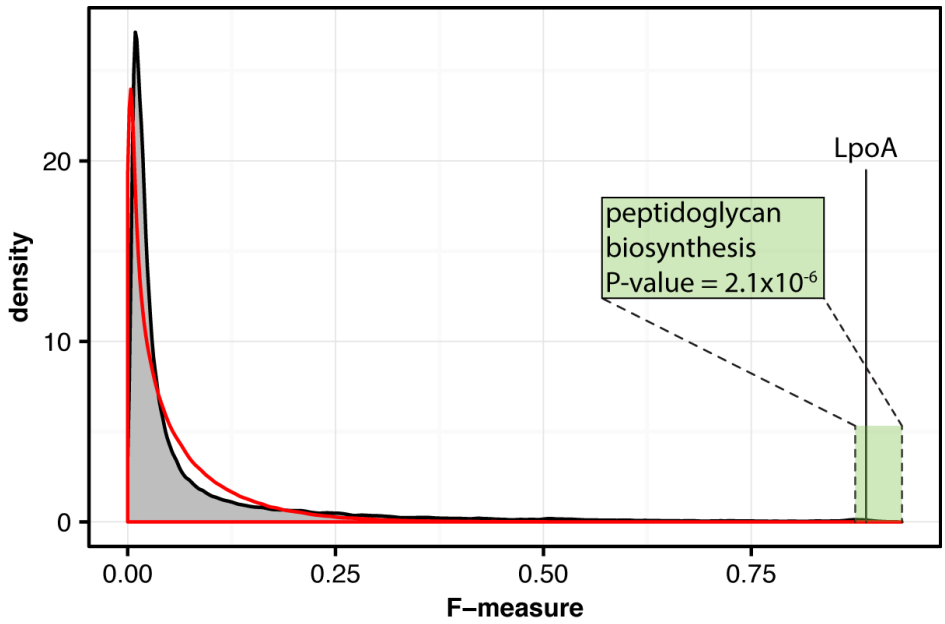

Figure 37: ODD Domain co-occurrence with all gene clusters in EggNOG, density distribution of F-measure, positive control LpoA gene cluster is indicated, most enriched GO term is peptidoglycan biosynthetic process. Red line is the beta distribution fit.

4.2.3 Incorporating phylogenetic tree comparison aids ranking of physical interaction partners

Co-occurrence measures alone were shown to rank the positive control proteins within the top 0.2%, and after filtering for localization within the first 10 hits. However, I wished to expand this approach to proteins for which there is no known interaction partner. Moreover, it became clear that comparing phylogenetic distributions from both domains and gene clusters is an imperfect process. This is largely because both domain and gene cluster phylogenetic distributions are defined through homology, where setting a cutoff will result in both false-positive and false-negative results. Hence I decided to refine this domain-protein detection method by incorporating information present in the sequences of both the domain and the protein.

It has been shown in the past that comparing the phylogenetic trees of one protein and comparing it to the corresponding tree of another protein can be a good indicator of protein-protein interaction. I adapted this method to protein domains, which to the best of my knowledge is the fist attempt to do this. Indeed comparing phylogenetic trees for over 100.000 gene clusters is extremely resource

intensive (see 4.3.1.4) and took several months of computing time on the available High Performance Computing infrastructure.

The first of a family of methods that compare phylogenetic trees is called MirrorTree (Pazos and Valencia, 2001). MirrorTree has the disadvantage that results have a high background (see 4.3.1.3), due to phylogenetic tree similarity to the underlying Tree of Life (TOL). I could also recapitulate this high background in my results, by observing the ranking of the positive control domain-protein pairs. In order to choose for the optimal phylogenetic tree comparison method I then corrected for the high background in the MirrorTree results as described in the TOL-MirrorTree method (Pazos et al., 2005). TOL-MirrorTree indeed improved the ranking of the positive control pairs compared to MirrorTree. However I wanted to compare my results to ContextMirror (Juan et al., 2008), which is the state-of-the-art method for removing background signal (see Materials and Methods section 6.3.11). In the interest of space I will only detail ContextMirror results here.

ContextMirror is even more resource-intensive since all phylogenetic trees in the analysis need to be pairwise compared, resulting in over 5 billion unique pairwise comparisons (see 4.3.1.4). Ranking the positive control pairs using the ContextMirror method typically gave results that match the predictive power of the co-occurrence methods alone.

More importantly, ContextMirror results hold information that is not readily captured by co-occurrence measures alone. Encouragingly, the positive control gene clusters are among the few gene clusters that rank very highly with both methods (see Figure 38). This observation lead me to think that incorporating information from both co-occurrence measures, and ContextMirror into a compound score would yield superior results than any of these methods alone. I experimented with different methods of incorporating the scores from the F-measure, the Mutual Information, and the ContextMirror methods in one score. The harmonic mean of all 3 scores, was chosen as it maximizes the ranking the positive control pairs. As an example, the ODD-LpoA gene pair ranks within the first 50 gene clusters using this compound score. This ranking result is among all gene clusters; this result can be further refined when limiting the search space, by making use of prior knowledge on possible interacting partners, such as their localization.

I also addressed using ContextMirror scores complementary to co-occurrence F-measure improves the enrichment for genes in the same process as the input

domain. Indeed, for the ODD domain, there are fewer than 170 gene clusters that rank within the first 500 gene clusters for both F-measure and ContextMirror (see Figure 38, top-right corner). Those gene clusters are over 12-fold enriched in peptidoglycan biosynthetic process genes, yielding a BH-corrected P-value of $2x10^{-8}$.

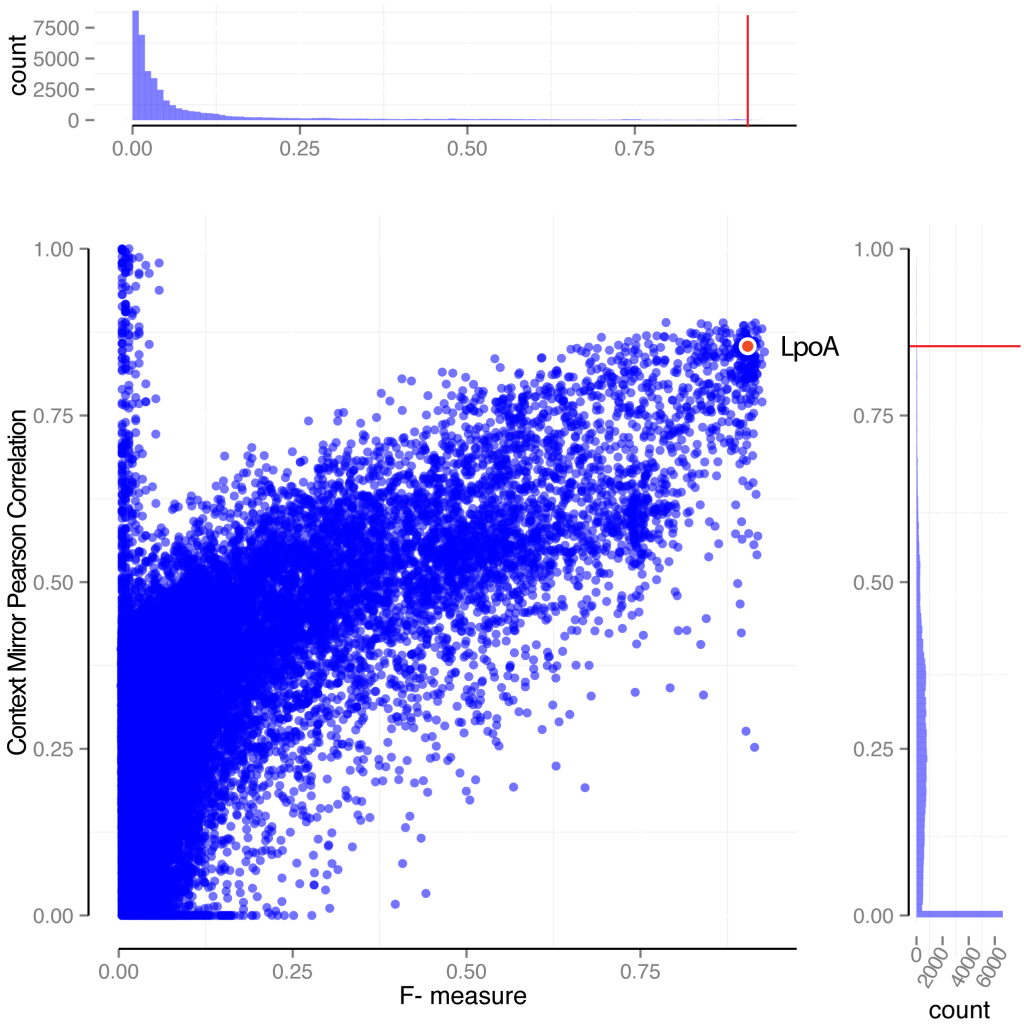

Figure 38: Context Mirror compared to co-occurrence F-measures when using the ODD domain as input. Each point corresponds to a gene cluster, LpoA gene cluster is indicated in the scatter plot and in both marginal histograms.

4.2.4 Exploring pipeline results provides hints for possible interaction partners

By exploring the coevolution of proteins from a diverse range of species chosen to span the bacterial kingdom (see 6.3.1), I was able to uncover potential

interacting partners of proteins involved in cell wall biosynthesis and degradation. Here I present a selection of these potential pairs of domain and interaction partner in order of phylogenetic similarity to the bacterium *E. coli*. Although these results imply interesting functional links among proteins involved in PG biosynthesis and turnover, at the time of writing these links were not yet experimentally validated (see also 4.3.2).

4.2.4.1 *Escherichia coli* amidase AmiD

Amidase enzymes are responsible for removing the peptide side-chain from the glycan strand, by cleaving the bond between the L-amino acid and the N-acetylmuramic acid. Thus amidase function is vital for recycling of cell wall components, as well as cleaving the septal cell wall during division.

In *E. coli*, four such enzymes are present in the periplasm. AmiA, -B, and –C are soluble proteins, while AmiD is anchored in the outer membrane (Pennartz et al., 2009). Using my pipeline, I defined a region of 64 amino-acids in the C-terminus of the AmiD (see Figure 39), which is conserved across species in the *Enterobacteriales* domain. This region, formerly present in the Pfam database as Pfam-B PB323, is also present in an amidase protein conserved among species in the Pseudomonales domain.

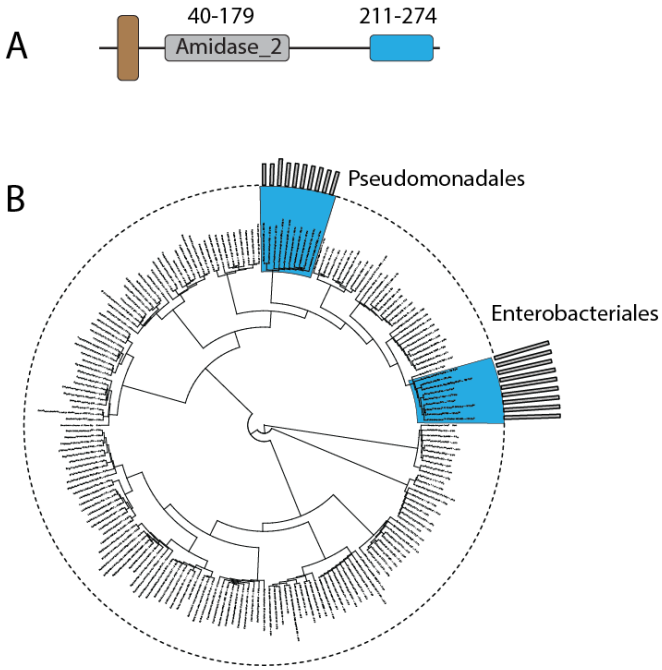

Figure 39: Amidase AmiD conserved region annotation in *E. coli*. 64 amino-acids in the C-terminal region are found to be conserved in amidases in *Enterobacteriales*, as well as in Pseudomonadales species

Among the first hits in proteins coevolving with this region are functionally related proteins, such as AmpD and lipoprotein NlpC. Similar to the protein holding the query conserved region, both AmpD and NlpC are involved in the recycling of cell wall components. On one hand, AmpD is also a N-acetylmuramyl-L-alanine amidase, cleaving peptides from the glycan strands. On the other hand the molecular role of lipoprotein NlpC is not yet uncovered in *E. coli*. NlpC features a P60 domain, which is shown to belong to a diverse family of domains that is even conserved among eukaryotic genomes (Anantharaman and Aravind, 2003). While the substrate of most types of P60 domains is still elusive, it is thought to also be involved in peptidoglycan degradation (Firczuk et al., 2007).

4.2.4.2 *Pseudomonas aeruginosa* endopeptidase MepM

MepM is an endopeptidase that cleaves the crosslink between the m-Dap and D-Ala residues on the peptide side chain. The endopeptidase activity of MepM has been associated to PG biosynthesis during cell growth (Singh et al., 2012). Studies involving radioactively labeled m-Dap residues in *E. coli* demonstrated that overexpression of any of the three *E. coli* PG endopeptidases (MepM, MepS and MepH) from a plasmid will rescue the lethal effect of a triple mutant (Singh et al., 2012). This suggests that these endopeptidases are functionally redundant. However, *E. coli* MemM is the only one of these three that belongs to the M23

family of peptidases, featuring both M23 and LysM cataltytic domains. *Pseudomonas aeruginosa* MepM does not feature a LysM domain but is annotated only with a M23 peptidase domain. Using techniques detailed in this thesis, I found that *P. aeruginosa* MepM contains a region that is conserved among *Pseudomonas,* as well as *Azotobacter*, suggesting it is conserved among *Pseudomonadaceae* (see Figure 40).

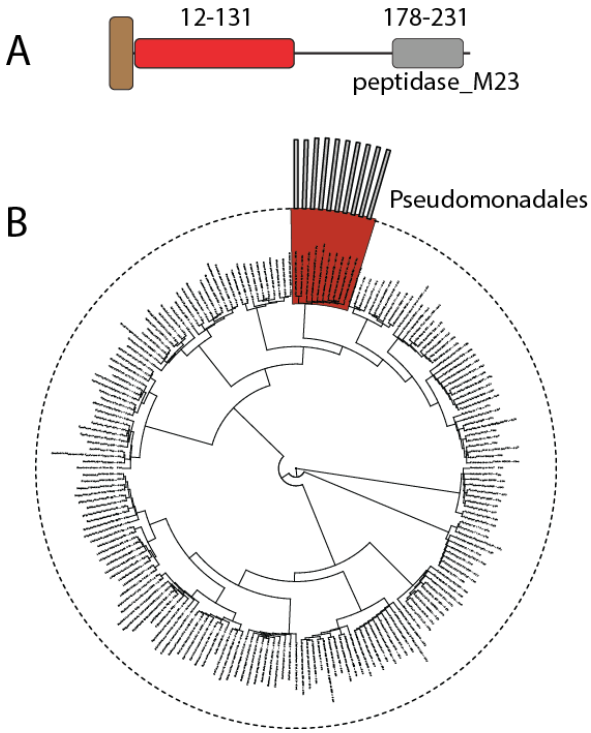

Figure 40: MepM DD-endopeptidase domain annotation in *P. aeruginosa* PAO1. Region shown in red is conserved among homologues of this protein in Pseudomonadales.

Interestingly, I found that this region appears to have co-evolved with a homologue of the NlpD lipoprotein (P45682). The *P. aeruginosa* NlpD features both M23 and lysM domains, and is thus identical in domain content to the *E. coli* NlpD. Moreover, the *E. coli* homologue has been shown to activate the function of another peptidoglycan hydrolase, amidase AmiC (Uehara et al., 2010). Our results imply that in *Pseudomonadaceae*, this regulation link is different, with the NlpD homologue activating the function of endopeptidase MepM. While both NlpD-regulated proteins, AmiC and (hypothetically in *P. aeruginosa*) MepM are involved in PG hydrolysis, their catalyzed reactions are different. My results imply that these regulation links among PG hydrolases changed within the class of Gammaproteobacteria, presenting an interesting hypothesis.

4.2.4.3 *Streptomyces coelicolor* transpeptidase SCO4013

E. coli protein PBP2 is an essential transpeptidase (TP) that is linked with cell elongation, and maintenance of cell shape (Banzhaf et al., 2012; Typas et al., 2012). The *Streptomyces coelicolor* homologue of PBP2 is protein *SCO4013*, which I found that features an 85 amino-acid domain conserved across all species in the Streptomycetales order. After analyzing the domain further, I found that it has likely coevolved with lipoprotein LpqB, itself featuring 2 domains that are known to be involved in sporulation-related signaling in *B. subtilis* (Setlow, 2003).

PG remodeling proteins such as hydrolases have been shown to play a role in sporulation both in B. subtilis and S. coelicolor (Haiser et al., 2009; Setlow, 2003). However, there is yet no evidence of the cross-talk with PG synthases during this process. By examining the coevolution between a domain conserved in a peptidoglycan synthase across *Streptomyces,* an interesting hypothesis forms also around the regulation of PG synthases during sporulation in *Streptomyces coelicolor* and related species.

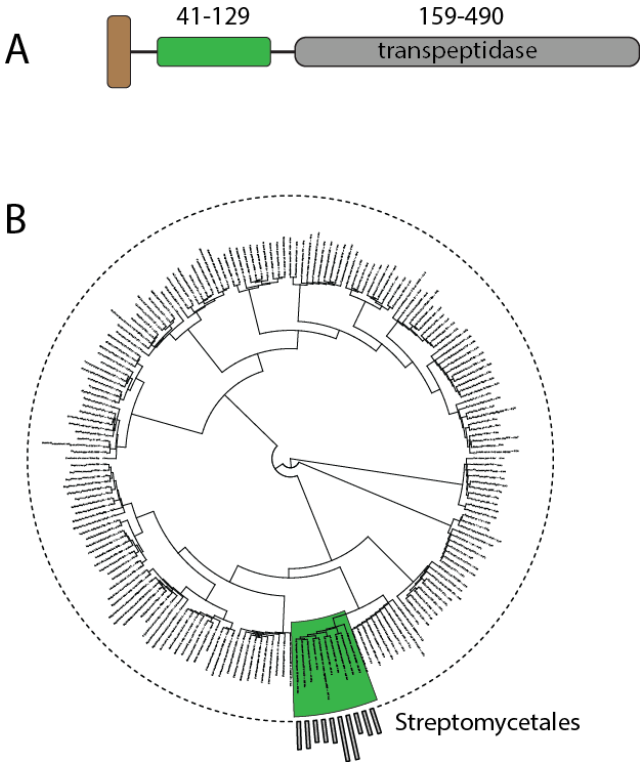

Figure 41: Domain annotation of *Streptomyces coelicolor* transpeptidase. Region shown in green is conserved among all homologues of this protein in *Streptomycetales*.

4.2.5 Refining the phylogenetic distribution of domains and regions in core process proteins

As an intermediate result of step 4 of the pipeline (see 4.2.1), I was able to refine the phylogenetic distribution for each of those domains and regions. For a few of those domains, their phylogenetic distribution came to stark contrast with prior knowledge. As an example, the ODD domain, was thought to be confined in the bifunctional PBPs of Gamma-proteobacteria. After exploring its phylogenetic distribution with the current pipeline, it became clear that the ODD domain spans more classes of bacteria, such as alpha-, beta-, and Gamma-proteobacteria. This result is recapitulated in the independent analysis performed by Ruth Eberhardt to annotate this domain in the Pfam database (see protein domain family PF17092 under http://pfam.xfam.org/family/PF17092, and Figure 43). Moreover, the UB2H domain can also be found outside of Enterobacteriales (see Figure 42).

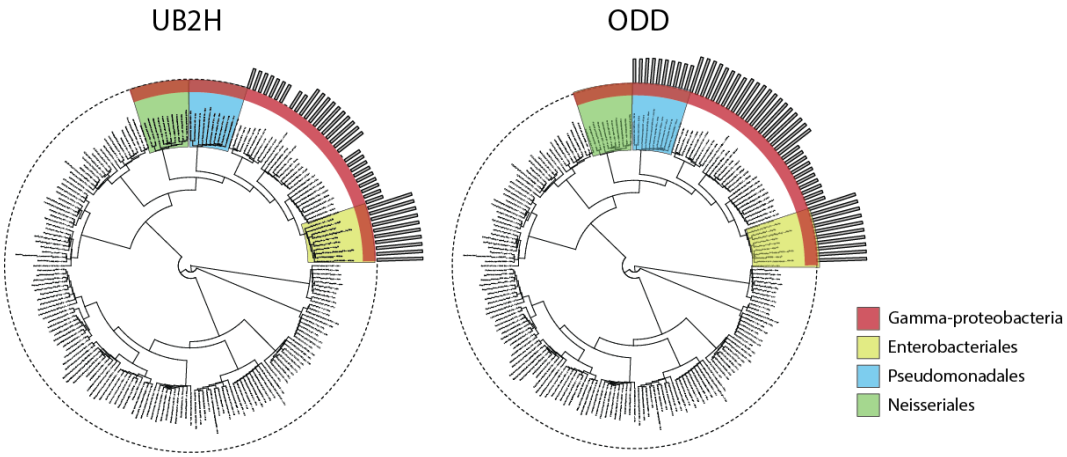

Figure 42: PBP1B UB2H domain and PBP1A ODD domain phylogenetic distributions. Species chosen for the phylogenetic tree are representatives across bacterial classes and orders.

4.2.6 Annotating PBP conserved regions as domains in Pfam

In the scope of this project, I collaborated with Ruth Eberhardt (Curator, Protein families, EMBL-EBI), Alex Bateman (Senior Team Leader, Protein sequence resources, EMBL-EBI) and Rob Finn (Team Leader, Protein Families, EMBL-EBI) to introduce into the Pfam database (Finn et al., 2014) non-annotated domain families with a focus in Penicillin-Binding-Proteins.

Together with Athanasios Typas, I identified a set of regions in modular proteins that were conserved across closely related species. Subsequently, I acquired their phylogenetic distribution using a part of the pipeline I developed (section 6.3.6 Phylogenetic distributions for PfamB domains and inter-domain regions). Examples include the ODD domain encoded in the *E. coli* mrcA (PBP1A) gene, identified at amino-acid positions 315-422, and a domain in the N-terminus of the *Caulobacter crescentus (strain CB15)* CC_3277 gene, which is also a class-A PBP.

Using these observations Ruth Eberhardt executed the Pfam domain analysis pipeline, and performed all necessary quality control checks. The domains described above are now part of the Pfam database as domain families PF17092 and PF17093.

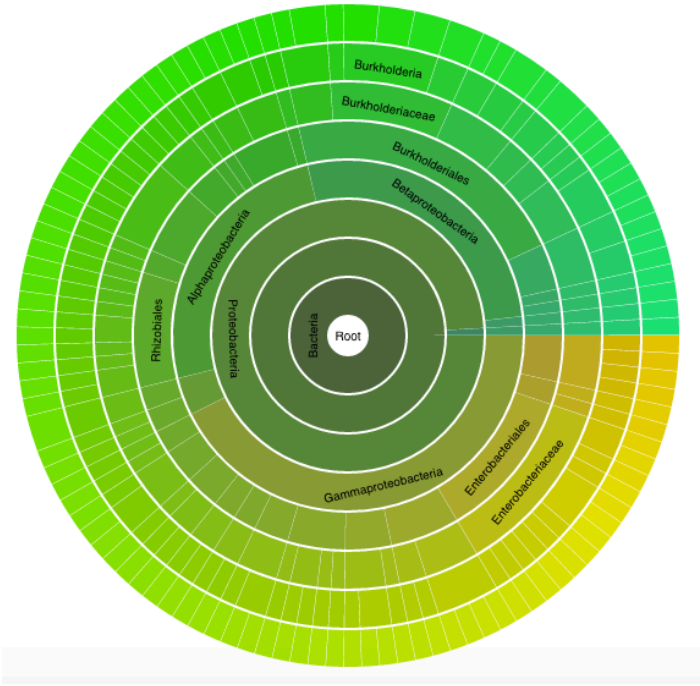

Figure 43: Phylogenetic distribution of ODD region, renamed as PCB_OB in the Pfam database (Finn et al., 2014)

4.3 Perspectives

In the present study, I have explored the modularity in the domain content of protein machineries in the core process of peptidoglycan cell wall biosynthesis and remodeling. My aim was to leverage the domain content variability of these proteins to detect their possible interaction partners.

The domain content of these proteins changes across evolution, and can vary also within species. Domain content has been shown to play a central role in the spatial and temporal regulation of these protein machineries in *E. coli* (Egan et al., 2014; Typas et al., 2010). This regulation was shown to take place through cofactor proteins that co-occur across species with certain domains. However, domain content across evolution has largely remained an unexplored source of information. I have shown that by using the co-evolutionary attributes of these domains one can narrow down their possible cofactor proteins.

Moreover, I analyzed the domain content of over 50 *E. coli* proteins, and that of their homologues across 20 species. After quality control, I focused on more than 75 domains and conserved regions that occur across species in proteins that belong to the core bacterial process of cell wall biosynthesis and remodeling. For each one of those domains I compared its phylogenetic characteristics to those of all gene clusters in the EggNOG database, acquiring ranked lists of possible cofactor proteins.

4.3.1 Overcoming technical challenges of domain-protein coevolution analysis

In order to identify potential interacting partners for the domains and regions I explored, I adapted methods that are widely used to detect whole protein coevolution. Here I discuss technical limitations of these methods and their adaptations, with a focus on their degree of computational complexity, as well as limitations with the underlying domain or protein detection across species.

4.3.1.1 Species selection: balancing diversity and coverage

A study observed that with increasing amount of sequences (400 at the time the study was conducted), there was no improvement in the quality of results of phylogenetic comparison methods (Sun et al., 2007). However, the authors concluded that selecting which species to include in a coevolution analysis will

heavily influence the results. In the same study they showed that selecting species according to specific guidelines for phylogenetic tree coverage and diversity improves co-occurrence analysis results.

Currently with more than 5000 fully sequenced bacteria (source: NCBI Genome, data as of May 2016), one needs to pay special attention to the species selection in order to satisfy the above guidelines. While selecting few sufficiently divergent species would fulfill the above guidelines, this would in fact prohibit the detection of conserved domain and regions that are evolutionary confined among closely related species.

In the current study I follow the species selection done in the EggNOG v4 database (Powell et al., 2014). Briefly, species diversity in EggNOG is ensured by clustering all available species according to their marker genes (Mende et al., 2013). Subsequently, representative species are selected from each cluster according to their genome assembly quality, as well as their utility as model species.

4.3.1.2 Limitations in defining domain and gene conservation

To detect domain-protein co-evolution, I aggregate the results of two approaches, which I adapted from the field of protein-protein coevolution: species co-occurrence and phylogenetic tree similarity (Juan et al., 2008). Both for co-occurrence across species, as well as for ContextMirror the first step is to acquire the set of protein homologues across species. In my implementation, adapted for domain-protein coevolution, I need to acquire the set of conserved domains as well as the set of protein homologues.

Although commonly used, acquiring this set of homologue proteins across species is often an imperfect process. Indeed, visualizing conservation across species makes clear that conservation is often a gradient as one moves away in phylogeny from the species used as query. Setting a threshold to define the set of species holding a protein homologue is thus likely to result in false positives or false negatives. This issue is especially evident when relying on the widely-used BLAST algorithm (Altschul et al., 1990). Algorithms such as jackHMMER (Johnson et al., 2010) partially mitigate the problem by re-defining the query HMM model of a given query sequence using its own search results. By iterating the steps of a) search and b) re-definition based on the search results, jackHMMER can achieve higher recall of protein or domain homologues across

species. However, a drawback that comes with this approach is that an HMM profile can diverge from the original query sequence.

For the definition of the phylogenetic distribution of a conserved region I used the jackHMMER algorithm. Diverging from the input query sequence is avoided in the current pipeline by setting a strict threshold in the jackHMMER settings. On the other hand, for defining the phylogenetic distribution of a domain in Pfam I used the existing HMM profile, calculated across the Pfam sequence base. In both cases, spurious hits are limited since I require the domain or region hit to be framed by the same catalytic domains as in the original protein used as query.

The set of domains I focused on is limited, which allows me to fine-tune the computational methods I use to get an optimal phylogenetic distribution. However, the set of proteins (corresponding to gene clusters) I search against is much larger and does not allow me to apply the same approach I used for the domains and conserved regions. For the phylogenetic distribution of proteins, I use the EggNOG v4 database (Powell et al., 2014). Genes are added in EggNOG gene clusters by reciprocal best Smith-Waterman match to at least 2 existing genes in the cluster (forming a triangle), in a fashion similar to the pre-existing COG database (Tatusov et al., 2000; Tatusov et al., 1997). This reciprocal best match requirement, as well as the triangle rule reduces the addition of spurious genes in a gene cluster.

However, a limitation in EggNOG is that it assumes that a protein is conserved in its entirety across species. This is rarely the case with modular proteins, which often loose or gain large parts across evolution. However advanced, EggNOG's clustering algorithm is not perfect. Indeed, exploring our results we often see cases where a gene cluster is either more strictly or more loosely defined. In the first case, the gene cluster is missing homologues of a gene, while in the latter case it contains genes that are not homologues of the seed gene. Regarding these issues, I am in contact with the main person maintaining the current iteration of the EggNOG database, Jaime Huerta-Cepas (Peer Bork group).

In conclusion, for phylogenetic distribution comparison, both domain (query) and protein (search space) phylogenetic distributions play an equally important role. The often-imperfect definition of a gene cluster thus leads to mismatches between a domain and a protein with little room for improvement if one is to use existing gene clusters to acquire protein phylogenetic distribution. Improving

gene cluster definition is beyond the scope of this study, however I am in contact with experts on this matter.

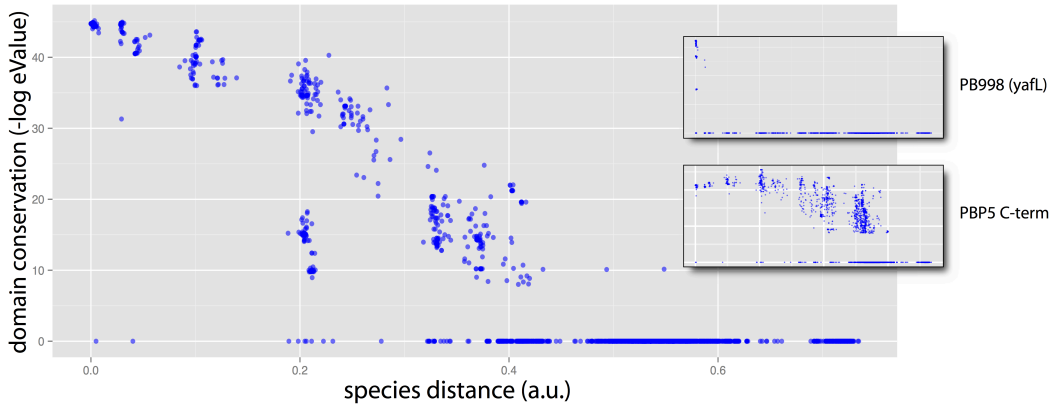

Figure 44: Visualization of protein sequence conservation across species. Main panel shows the similarity of the ODD region across species compared to species distance (Mende et al., 2013), origin (left-most) species is *E. coli* strain K12. Inlets show PfamB domain PB998 confined only in *Escherichia*, *Shingella*, and *Citrobacter* (top); In contrast, PBP5 C-terminus domain is widely conserved among bacteria (bottom)

4.3.1.3 Limitations of phylogenetic tree comparison methods

Phylogenetic tree comparison methods largely rely on pairwise distances between protein homologues, calculated via a multiple sequence alignment (MSA) algorithm. Although not immediately obvious, this makes tree comparison methods partially subject to the limitations of defining protein conservation (see section 4.3.1.2 Limitations in defining domain and gene conservation), since the MSA will be calculated among the homologues detected by a sequence similarity method.

However, the greatest drawback of phylogenetic tree comparison methods is the high degree of similarity between the phylogenetic trees of any two proteins (Ochoa and Pazos, 2014). This background similarity stems from coordinated changes across the two proteins due to speciation events, and not because of the effects of coevolution. Since both proteins are similarly affected by the underlying speciation process, any two protein trees will exhibit a certain degree of similarity.

Several correction methods have been proposed to mitigate this issue (Pazos et al., 2005; Sato et al., 2005). Exploring different phylogenetic tree comparison methods, I have found the ContextMirror method (Juan et al., 2008) to be

superior in alleviating the background speciation events. ContextMirror builds on top of MirrorTree's output, and defines the co-evolution profile of a protein family as a vector of MirrorTree results across all other proteins families. Subsequently, the similarity between pairs of co-evolution profiles is calculated by means of their Pearson's correlation coefficient. Although ContextMirror results are of much better quality, the required profile generation steps significantly increase computational time complexity.

Predicting domain-protein interaction partners is expected to have limited predictive power, owing to limitations in precise definition of protein phylogenetic distribution. In fact, the fragmentation in gene cluster definition in different phylogenetic tree levels (Powell et al., 2014) results in a large number of gene clusters that further dilute our signal (see section 4.3.1.2 Limitations in defining domain and gene conservation).

Moreover, most phylogenetic tree comparison methods show clear limitations owing to the high background similarity of any two phylogenetic trees. As discussed above speciation events drive this background similarity and various tree comparison methods mitigate this effect to various degrees (Juan et al., 2008; Pazos et al., 2005; Pazos and Valencia, 2001).

In my results I demonstrate that by using these two approaches in tandem, their individual limitations can be partially overcome. Indeed by combining results from both co-occurrence and a tree comparison method (Juan et al., 2008), I observe an improved ranking for our positive controls, as well increased enrichment for proteins in the same process.

4.3.1.4 Reducing the computational complexity of phylogenetic tree comparison

Methods to detect protein coevolution tend to be quite computationally intensive. Methods such as MirrorTree rely on Multiple Sequence Alignment (MSA) software such as ClustalW to generate phylogenetic trees. ClustalW has a complexity of $O(S^2)$, where S is the number of sequences, or in this case the number of species in the analysis. Even using the latest MSA algorithms, such as Clustal Omega, the complexity is dominated by the comparison of all pairwise species distances, which is essentially a Pearson correlation. Assuming the 3-way

Toom–Cook algorithm is used to calculate products, the total time complexity of MirrorTree will amount to $O(S^{1.465})$ [4].

In the context of the current study, I took advantage of the high-performance computing infrastructure available and calculated ContextMirror phylogenetic profiles for all gene clusters in the EggNOG database. ContextMirror has the advantage of being less influenced by the background speciation additional. However, ContextMirror adds the step of calculating phylogenetic profiles, by essentially pairwise comparing phylogenetic trees of all genes. This scales quadratically with the number of genes (G) compared, resulting in a time complexity of $O(G^2 \times S^{1.465})$.

The number of genes *G* vastly exceeds that of sequenced species *S* (G>>S), making a pairwise protein phylogenetic tree comparison practically impossible within a reasonable time-frame even with use of high-performance computing equipment. Taking this a step further, each gene corresponds to a protein that potentially features a number of domains (D, where D>G). Given the more-than-cubic time complexity of these methods, a pairwise domain phylogenetic tree comparison would take several times the computational time of the protein-level comparison.

I conclude that, despite the potentially high value, domain-domain coevolution has not been explored systematically for a number of years (Ochoa and Pazos, 2014). A probable reason is the increase in the number of available sequences, alongside the poor scalability of available methods with increasing number of sequences (Ochoa and Pazos, 2014). In the current study I circumvent this limitation, by focusing on domains and conserved regions that occur in specific proteins.

In the current implementation I tackled these issues by first selecting a set of widely conserved gene clusters and used them as reference against which I calculated phylogenetic profiles. This reference set (*R*) of 5665 gene clusters

[4] Since additions and subtractions happen in constant time, the Pearson correlation time complexity is dominated by the square root calculation algorithm. Square root calculation complexity is in turn dominated by the multiplication complexity. The efficient and widely-used Toom-Cook algorithm has a time complexity of $O(S^{1.465})$ (Knuth, D.E. (1997). Art of Computer Programming, Volume 1: Fundamental Algorithms. Journal of the American Statistical Association *1*, 435-455.

corresponds to the species-wide calculated COG-level group of gene clusters in EggNOG (see also section 6.3.3 Protein phylogenetic distributions). Thus the complexity of calculating phylogenetic profiles is reduced to $O(G \times S^{1.465})$ [5], as compared to $O(G^2 \times S^{1.465})$ of the original ContextMirror method.

The second step of ContextMirror would still dominate the complexity of this analysis since it is $O(G^{3.465})$. However, this second step is not required and thus omitted in this analysis. Instead I compare the phylogenetic profiles of all gene clusters to the phylogenetic profile of a domain of interest, which is sufficient to detect domain-protein coevolution.

In total, I adapted the ContextMirror algorithm for this application, at the same time drastically reducing its time complexity. Importantly, this optimization for running time still provides very good results, as demonstrated by the very high ranking of our positive control proteins.

4.3.1.5 Modularity within cofactor proteins

While modularity within large protein machineries such as the bifunctional PBPs is clear, cofactor proteins were up to now discussed as if they are contiguous entities. This is not always the case however. For example, the large (678 amino acid) cofactor protein LpoA is annotated in the Pfam database to have two LppC domains.

Importantly, these two domains of LpoA were shown to have different function in the cell (Gray et al., 2015). The C-terminus domain of LpoA is sufficient to activate the function of PBP1A, while the remaining N-terminus was found to contain a tetratricopeptide repeat (TPR) region. TPR domains are present in a wide range of proteins across evolution and are often used to mediate protein-protein interactions (D'Andrea and Regan, 2003; Goebl and Yanagida, 1991; Zeytuni et al., 2012). It was further demonstrated that the TPR N-terminus domain in LpoA compensates loss of CpoB, another cofactor protein that was shown in the same study to interact both with PBP1B and with the Tol machinery to coordinate peptidoglycan synthesis at the division site. Interestingly, CpoB also contains TPR domains that span the C-terminal half of the protein.

[5] Adapted ContextMirror complexity is $O(R \times G \times S^{1.465}) \equiv O(G \times S^{1.465})$, since R is a fixed number equal to 5665.

These results highlight the importance and unexplored versatility of modular cofactor proteins. A limitation of the present study is that due to resource constraints I did not address cofactor protein modularity. Although some cofactor proteins, like LpoA are shown to have modular functions, exploring coevolution between a domain and all modules of all potential cofactor proteins quickly scales to become prohibitively computationally expensive (see 4.3.1.4).

4.3.1.6 Detecting interconnections mediated by broadly-used domains

In recent studies, peptidoglycan biosynthesis enzymes have been shown to directly interact both in vivo and in vitro (Banzhaf et al., 2012; Bertsche et al., 2006). These studies showed that PBP1A, and PBP1B interact respectively with monofunctional transpeptidases PBP2, and PBP3. In both interaction cases, the monofunctional transpeptidases feature N-termini domains called "PBP dimer domain". Although the function of this domain has not been precisely defined, they are believed to mediate PBP polymerization.

Since both PBP2 and PBP3 use the same domain to bind to their interaction partners, the specificity of these domains cannot be resolved at the domain level. Moreover, the *PBP dimer* domain is widely conserved across bacterial species, thus failing both criteria required for their coevolution analysis using this pipeline. Similar limitations exist with the aforementioned TPR domain, which have also been shown to mediate protein-protein interactions across all kingdoms of life, yet computational prediction of TPR binding partner of substrate has yet to be demonstrated (Zeytuni et al., 2012).

4.3.1.7 Combining known experimental and genomic context information

The results of the current pipeline for every domain are presented as a list of ranked potential interaction partner proteins. As indicated by the results, the current pipeline successfully enriches for proteins that are functionally related to the query domain. However, information such as genomic context, experimentally shown interactions, as well as text-mining can be instrumental to further narrow down to a potential physical interaction partner. Thus I provide all such information for every possible interaction partner.

To do so, I query the STRING database for such information between the predicted interacting partner protein, and the protein holding the input domain in

the same species. This feature is expected to aid the non-expert user make immediate use of the results. Moreover I expect that this will save the user a significant amount of time, as compared to having to manually search STRING or other databases.

4.3.2 Outlook

Experimental validation

At the time of writing the present thesis, I had shown the potential of this coevolution pipeline in retrieving proteins known to be interacting to specific domains in peptidoglycan synthases in *E. coli*. Naturally, the next step in the current project is to experimentally verify the predictions of this pipeline for domains with no known binding partner. This would serve the dual purpose of proving the relevance and predictive power of this pipeline, as well as helping to fine-tune the parameters of the pipeline on a new set of domain-protein interaction pairs.

The number of experiments scales quickly with the number of tested protein-domain pairs. Indeed validating the results of this pipeline will result in a medium-scale protein-domain interaction screen. Testing possible interacting partners of the 48 PfamA and PfamB domains, will result in an estimated 500 domain-protein pairs. Furthermore, the domains and proteins to test belong to a diverse range of species, for which there are often no advanced genetics tools available. Taking the above into account, I decided to use the yeast-two-hybrid (Y2H) method as the screening method of choice.

The main advantage of Y2H for this application is that it is easily scalable, and several of its steps can be readily parallelized. Moreover, beyond the extraction of genomic DNA for domain and proteins, the species tested is not expected to influence the efficiency of this screen.

Using the described coevolution pipeline, I have also acquired potential interacting partners for domains that are specific to species such as *P. aeruginosa*, *V. cholera*, *S. pneumonia*, and *B. subtilis*. In parallel to validating such hits using the Y2H method, it is in my immediate plans to communicate such hits to collaborating labs that specialize in these species to aid me in potential follow-up experiments.

Future directions

Validating the results of the current pipeline will demonstrate the relevance and predictive power of this pipeline in detecting interaction partners of cell-wall related proteins. Predicting physical interactions in the periplasmic space has the added advantage that it would be the principal way to regulate peripalsmic processes, owing to lack of ATP. However, I envisage the application of this pipeline to other core processes featuring highly modular proteins with the aim to uncover their potential interaction partners.

Moreover, the increase in domain-protein pairs that are shown to interact will serve to fine-tune the parameters of this pipeline. In fact the co-occurrence and phylogenetic tree comparison results are now equally weighed in defining the ranking the final hits. With enough known protein-domain interaction partners, I can fine-tune the contribution of each component on the final score. Last but not least, with a sufficient number of known domain-protein pairs, it is straightforward to apply a machine learning approach to optimize each component's effect on the final score.

4.3.3 Conclusions

Exploring domain-protein coevolution can be used to detect interaction pairs

An increasing amount of evidence shows that protein domain content plays an instrumental role in the regulation of core processes, such as the peptidoglycan cell wall biosynthesis. In the present study I showed that using the domain content of peptidoglycan biosynthesis machineries such as the bifunctional PBPs one can narrow down to their potential protein interaction partners. In fact, these interaction partners can be instrumental in uncovering new regulatory connections, since post-tranlational regulation in the periplasmic space –in the absence of phosphorylation, or other post-translational modifications- would be principally conducted through physical interactions.

Moreover, I showed that domain-protein co-occurrence and phylogenetic tree similarity hold complementing information that can further aid in the ranking of such potential interaction partners. Lastly, I showed that the use of prior localization information is sufficient filtering to list known domain-protein interacting pairs within the top-10 pairs.

Detecting potential cofactors for cell wall protein domains across evolution

After demonstrating the potential of this predictive pipeline, I analyzed the domain content of more than 50 cell wall related proteins and their homologues across 20 diverse species. For each such domain I have produced a ranking of possible interaction partners, supplemented with localization, and genomic context information.

To experimentally validate the hits for the domains that are also conserved in *E. coli* proteins, I am currently collaborating with colleagues that undertake a Y2H screening approach to test highly ranked protein-domain interactions.

4.4 Contributions

The present study was conceived and performed by myself. Incorporation of the ODD and other domains to the Pfam database was done by myself and Ruth Eberhardt (Finn team, EMBL-EBI). Yeast-two-hybrid validation of domain-protein pairs will be performed by Matylda Zietek (Typas Lab). I wish to thank Sonja Blasche for her help in establishing the Y2H screen. I wish to thank Rob Finn and Alex Bateman for useful discussions. I also wish to thank David Ochoa for useful discussion on phylogenetic tree comparison methods. Finally, I wish to thank Jaime Huerta-Cepas for useful discussion on methods used to construct the EggNOG database.

5 Discussion and concluding remarks

This chapter aims to outline the main contributions of my work, as well as place them on the landscape of the adjacent research fields. For an extensive discussion on the advances and limitations of the work presented in this thesis, I direct the viewer to the discussion paragraphs of each chapter: 4.3, 3.3, and 2.5.

In the field of molecular biology, our understanding of gene function is still lagging behind our ability to map new genes by sequencing different organisms. The advent of high-throughput reverse genetics approaches has helped narrow this gap, accelerating our ability to delineate gene function. Such approaches typically query the phenotype of a known genetic perturbation, often combining that with other genetic or chemical perturbations. Gene-gene or gene-drug interaction screens were first developed for the model fungus *Saccharomyces cerevisiae* (Parsons et al., 2006; Roemer et al., 2012; Schuldiner et al., 2005; Schuldiner et al., 2006; Tong et al., 2004) and later also applied to bacterial species (Brochado and Typas, 2013; Deutschbauer et al., 2011; Nichols et al., 2011; Pasquina et al., 2016; Phillips et al., 2011; Typas et al., 2008).

These approaches are mostly applied to microbes by measuring macroscopic level phenotypes, such as growth on solid surfaces, which serves as a proxy of growth fitness. Perturbation of several processes and pathways could lead to a growth fitness phenotype. Also measuring one simple phenotype allowed for highly increased assay throughput, facilitating the simultaneous expansion of such assays across chemical conditions and genetic perturbations.

This increased throughput allowed the generation of genome-wide genetic interaction screens, such as the ones performed in *Saccharomyces cerevisiae*, and *Schizosaccharomyces pombe* (Costanzo et al., 2010; Frost et al., 2012). These screens explored the underlying interconnections of genes, and also addressed the conservation or repurposing of such interconnections. In parallel, chemical large-scale genomics screens in those species allowed for querying and comparing mutant reactions across chemical stresses (Kapitzky et al., 2010). This allowed the discovery of drug mode-of-action conservation, and the identification of a novel small molecule that acts as DNA damaging agent across eukaryotic cells. While

quite successful in uncovering gene function and interconnections, the information content of such screens is limited to the measured phenotype.

On the opposite end of high-throughput phenotypic approaches, high-content screening (HCS) methods rely on the acquisition of hundreds of features of cells under chemical or genetic perturbation (Conrad and Gerlich, 2010; Zanella et al., 2010). These features are typically acquired from microscopy images through a feature extraction process. While rich in information content, these screens are harder to scale, and large number of acquired features is usually discarded to keep only those informative of phenotypes of interest (Singh et al., 2014).

Targeted phenotype acquisition holds a middle ground between measuring a single phenotype (such as growth), and measuring hundreds biology-uncoupled image traits (as in HCS methods). Directly measuring biologically relevant traits bypasses the often-complex steps of high-content screening. This allows for high throughput measurements of traits holding information that is often orthogonal to that of growth fitness.

In a recent application of such an approach, *B. subtilis* transposon mutants were used to uncover a set of genes unable to perform sporulation. Knockout mutants of these genes were then used to uncover the perturbed sporulation stage using fluorescence microscopy (Meeske et al., 2016). In a similar approach, a *P. aeruginosa* transposon (Tn) mutant library was made based on a background strain that features augmented biofilm formation (Cabeen et al., 2016). Visually inspecting mutant colonies then lead to the identification of *PA14_16550* and *PA14_69700* loci, Tn insertion or deletion of which abolished or augmented biofilm formation respectively. In an earlier application, visually exploring the morphology of *C. albicans* colonies allowed for the identification of genes involved in invasive growth, and biofilm formation (Ryan et al., 2012).

Microbial biofilms are of intense practical interest, since they allow species such as the opportunistic pathogen *P. aeruginosa* to survive in medical equipment. On the other hand, bacteria such as *B. subtilis* and *Myxococcus xanthus* can form spores that are impervious to environmental insults.

Despite the high practical value of knowledge surrounding the processes of biofilm and sporulation, high-throughput approaches targeting such phenotypes are scarce. Such a discrepancy could partially be owed to the fact that such screens

have no automated phenotype acquisition part, making the analysis of these screens extremely resource-intensive and thus hard to scale.

In the course of my work, I developed an image analysis software called Iris that can automatically quantify a multitude of microbial macroscopic phenotypes, including biofilm formation and sporulation. Importantly, this fully automated method can be applied at scale, across large numbers of mutant strains, and also across different growth conditions. Iris is currently used in the lab to quantify biofilm formation, as well as growth fitness phenotypes in two independent large-scale chemical genomics screens involving mutant libraries of *P. aeruginosa* (Liberati et al., 2006), and *E. coli* (Baba et al., 2006).

Both chemical genomics screens resulted in a wealth of data, often leading to yet-uncovered aspects of biofilm formation and maturation, as well as insights on other processes in both species. Results led to biological follow-up studies that are currently being conducted, and two manuscripts are in preparation in which I am also co-author. The potential of such large-scale targeted-phenotype screening methods is exemplified in the manuscript in preparation, which is available to the reader (see supplement).

Combining measurements of diverse phenotypes can bring added value to the data. For example congo-red biofilm staining and biofilm morphology characterization in colonies of *S. enterica* was shown to disentangle genetic components of different molecular pathways leading to the mature biofilm. Moreover, such added value can be obtained in the future, by combining automated phenotype acquisition with existing observations from previous large-scale reverse genetic screens. In an example of such an integration of newly acquired with older data, Shiver and colleagues integrated growth fitness data from a recent large-scale chemical genomics screen, with a smaller screen targeted on specific chemicals. This integration helped in elucidating the import mechanism of two neglected antibiotics and is detailed in a recent publication in which I am also co-author (Shiver et al., 2016).

More importantly, the Iris image analysis software is designed for easy extensibility to allow for an ever-expanding range of microbial phenotypes that can be quantified in a high-throughput way. The range of such phenotypes includes biofilm formation, microbial colony morphology, and sporulation. The potential of applications for each such phenotype acquisition are presented in the

manuscript submitted for publication (see also supplement of present thesis, as well as related chapter).

One such application of Iris in a targeted phenotype chemical genomics screen involved the detection of a chromophore reaction used to identify mutants involved in cell envelope biogenesis in the Gram-negative bacterium *E. coli*. Gram-negative bacterial cell envelope forms both a structural, and a permeability barrier that several antibiotic compounds cannot penetrate to reach their molecular targets. Although the object of intense study, envelope biogenesis and in particular the biogenesis of the outer membrane permeability barrier, and of the cell wall structural barrier are yet to be fully elucidated (Bos et al., 2007; Nikaido, 2003; Silhavy et al., 2006).

By means of establishing a new assay and using it in a small screen, Paradis-Bleau and colleagues identified mutant colonies involved in envelope biosynthesis across 4 growth conditions (Paradis-Bleau et al., 2014). This approach identified several genes involved in cell envelope biosynthesis, both recapitulating existing knowledge, as well as elucidating novel genes involved this process. One such novel factor was shown to be involved in the biosynthesis of a precursor used in the structural component of the cell envelope (Paradis-Bleau et al., 2014). Demonstrating the power of targeted-phenotype chemical genetic screens, this approach identified cell envelope defects for mutants of 70 genes, which had no growth phenotype when assayed in more than 320 conditions (Nichols et al., 2011).

To further systematically explore the cell envelope, I improved upon this screening method, and mitigated its limitations to acquire quantitative envelope defect measurements. Importantly, such quantitative measurements can then serve as a second layer of data analysis of a future chemical genomics screen, in order to compare phenotypic signatures of mutants across conditions (see corresponding chapter and discussion therein).

While the outer membrane of the Gram-negative bacterium serves as the permeability barrier, the peptidoglycan (PG) cell wall forms a rigid structural barrier protecting the cell against turgor, as well as environment osmolarity changes. PG is ubiquitous among bacteria, but also confined in this kingdom of life. This makes PG synthases, proteins called PBPs, the target of many beta-lactam antibiotics, such as penicillin.

Being the targets of penicillin, the molecular roles and biochemical functions of PBPs have been long elucidated. Yet, how these enzymes work as part of bigger complexes, how they are regulated, and how they interconnect with other processes in the bacterial cell envelope is a field we're only beginning to understand.

Recent evidence suggested that two *E. coli* PBPs are regulated by outer-membrane lipoproteins (Paradis-Bleau et al., 2010; Typas et al., 2010). Additionally, certain protein domains in these PBPs were shown to have coevolved with their cognate lipoproteins (Typas et al., 2010), providing hints for the niche-specific regulation of this process. In the absence of ATP, regulatory links in the periplasmic space of Gram-negative bacteria have to be mediated through physical protein-protein interactions.

Conserved interactions across evolution leave traces of coevolution among interacting partners; such traces can then be found by analyzing the gene sequences and genomic context across species. Aiming to uncover such regulatory relationships that include domains of proteins involved in PG biosynthesis and remodeling, I developed an approach that detects traces of coevolution, indicative of protein-domain interactions. I demonstrated that this method detects known domain-protein interacting pairs and can readily be used to uncover such possible interactions across bacterial species.

While similar approaches have been very successful in identifying coevolution of interacting protein pairs (Pazos et al., 2008; Szklarczyk et al., 2014), such approaches have never been adapted to protein domains. Importantly, such coevolution approaches turn the overwhelming volume of bacterial sequence information into a two-faced advantage. First, the amount of sequenced bacterial sequences allows for increased statistical power in observing traces of coevolution. Second, the limited number of genes in a species pan-genome (present in all strains of a bacterial species) helps to narrow down potentially co-conserved interacting partners (Donati et al., 2010; Jacobsen et al., 2011).

At the same time, colleagues have undertaken two parallel efforts to systematically elucidate such regulatory links in *E. coli*. The first such approach entails creating multiple-gene deletion mutants among genes involved in PG biosynthesis and remodeling to uncover possible regulatory relationships among the overlapping functions of these proteins. This unprecedented approach to systematically query the genetic interaction landscape among such proteins is further combined with perturbation of all mutants by a chemical stress. Chemical

compounds are chosen for their known affinity to specific players of PG biogenesis, thus shedding light into phenotypes indicative of mis-regulation upon perturbation.

The second approach also ongoing by colleagues in the lab is to uncover interactions between proteins involved in PG biosynthesis and remodeling by means of an affinity-purification approach, followed by mass-spectrometry (AP-MS). Efforts in this approach are still ongoing, however early results demonstrate the power of the approach by recapitulating most known interactions among cell wall related proteins and many new. Thus the computational method I developed helps to prioritize the set of interactions to be tested, or pinpoint the protein domain that is likely to mediate an experimentally observed interaction. This approach will also serve in the future to expand the knowledge gained from these experiments into possible regulatory connections that are conserved across evolution through the domain content of modular proteins.

In conclusion, during my work I helped expand the phenotypic landscape of multiple microbial species, by developing a software that can quantify diverse phenotypes in a high-throughput setting. I also made this software freely available for everyone to use and ensured that the freely available source code is easy to adjust to one's needs and to new assays so that it keeps serving the community in the future.

I then transitioned from the high-throughput data of such a screen to the mechanistic understanding of the Gram-negative cell envelope biosynthesis process and shed light in novel aspects of this process by building upon and improving this screening method. Moreover, I focused on the structural component of the cell envelope, the cell wall, to elucidate regulatory relationships mediated by protein domains. To this aim, I developed a molecular co-evolution computational approach that can readily be used to capture such relationships across evolution.

To summarize, in the course of my work I developed and applied high-throughput phenotypic methods, and molecular co-evolution methods to target the Gram-negative bacterial cell envelope.

6 Materials and Methods

6.1 Image analysis methods for automated microbial colony phenotyping

6.1.1 Software design

Iris is designed in a modular fashion to allow for easy extensibility to new assays and readouts. Each image-processing task described below is performed by a separate module. At the same time, different modules performing the same task are easily interchangeable. This allows for expert users to write custom made modules to fit the needs of assays not covered already by the distributed Iris version.

6.1.2 Picture processing

High throughput phenotypic assay quantification starts with a typically high-resolution picture of a colony array on a rectangular agar plate (see Figure 2). The first processing step for the software is to automatically rotate the picture so that the colony array is perfectly horizontal. Iris performs accurate image rotation using a technique widely used in OCR (Optical Character Recognition). This involves rotating the picture in 0.5° increments, and calculating the per-row sum of pixel brightness. The rotation in which the variance of the brightness sums is maximized is the one where colony rows are perfectly horizontal.

In the next step Iris detects and crops the plate boundaries. Plastic plate boundaries diffract light towards the camera upon side lighting. Iris detects the elevated brightness level of the plastic plate borders, thus detecting the plate boundaries.

A cropped picture containing only the colony array is then segmented into picture tiles, each holding only one colony. Since colonies are usually brighter than the background, summing brightness per-row and per-column will give brightness sum valleys and peaks. Peaks correspond to colony row or column centers, while local brightness sum minima correspond to inter-colony space (see Figure 2). Iris crops the picture in this inter-colony space, effectively segmenting

the picture into single-colony pictures, called tiles. As some colonies can generally grow larger than their holding tile picture, Iris implements an improvement of this algorithm, whereby tile bounds are allowed to vary to always accommodate an entire colony.

Every tile is then separately processed by one or several tile processor modules, each specialized quantifying a specific phenotype (e.g. sporulation). This design allows for independent colony quantification, but also easy incorporation of new readouts.

6.1.3 Phenotype quantification

Following image segmentation, each picture segment holds one colony. This image segment is called a tile and is independently processed to yield different colony phenotypes.

6.1.3.1 Colony bounds detection

The first step in any colony phenotype quantification is to accurately detect the colony bounds. Colony shapes, sizes, colors, as well as background vary substantially across the diverse applications of Iris. Subsequently, Iris uses distinct state-of-the-art techniques in edge detection to robustly detect colony bounds across different assays.

For applications where colonies are brighter than the background, Iris applies image thresholding algorithms, such as the Otsu (Otsu, 1979) algorithm. Typically such thresholding algorithms operate on the histogram of picture brightness, and attempt to select a threshold best separating the foreground (bright) from background pixels.

For applications where the brightness of a colony relative to its background is uncertain, or may vary within the assay, Iris employs the Marr-Hildreth algorithm (Marr A N and Hildreth, 1980), also known as Laplacian of Gaussian algorithm. This algorithm first applies a smoothing Gaussian filter to the picture. Subsequently a second order derivative of the Gaussian is calculated, whereby zero values denote sharp changes in brightness. These pixel locations are then used as colony bounds.

Colonies of *C. albicans* also extend into the agar, which is readily observable at the picture of a colony array. By applying two sequential image thresholding

algorithms of varying sensitivity, Iris captures both the extend of the filamentous agar invasion, as well as the (over-agar) colony size.

6.1.3.2 Colony size and opacity

Colony size is measured by several available software to be used as a proxy of growth fitness. Such measurement is performed by simply counting the number of pixels within the colony bounds.

Colony opacity is measured by summing the per-pixel over-background brightness values for all the pixels within the colony bounds. The over-background brightness for every pixel is in turn calculated by subtracting the pixel brightness to the average brightness of background pixels.

6.1.3.3 Sporulation

B. subtilis sporulating cells turn dark brown in minimal media. This pigmentation change takes place in the colony center. To robustly detect the colony center, Iris takes advantage of the fact the colonies were pinned in a square array. In this way, the X-axis displacement will be the same for all colonies in a column; the same is valid for the Y-axis displacement for all colonies in the same row. The X-axis displacement for colony centers in every column is then calculated as the median of all X-axis displacements in this column. Colony center Y-axis displacements are calculated in a similar fashion for each row, resulting in robust coordinate calculation for all colony centers.

To quantify the pigmentation change in the colony center, Iris assesses the three primary color channels in the cubic RGB representation per pixel. To do so, green and blue channel intensities are added together, and multiplied by a gain factor; the same process is done to the red channel. Subsequently the difference of the red channel product is subtracted from that of the green and blue channels. Since cells turn dark brown, pigmentation change towards brown color will only detect part of the change. I thus incorporated pixel brightness as part of the sporulation score formula:

$$color\ score = g_{RGB} \cdot [g_R \cdot R - g_{GB} \cdot (G+B)] + g_{SD} \cdot (S+D)$$
$$D = 255 - brightness$$

6.1.3.4 Biofilm formation

Colonies that form biofilm show increased dye binding, thus turning dark red. Iris detects this color change in a fashion similar to the detection of sporulating cells in *B. subtilis* (see 6.1.3.3). Briefly, Iris assesses the three primary color channels in the cubic RGB representation per pixel. Each color channel, as well as pixel brightness and color saturation are multiplied by a gain factor. Subsequently, the color score for each pixel is calculated as in the formula below:

$$color\ score = g_{RGB} \cdot [g_R \cdot R - g_{GB} \cdot (G+B)] + g_{SD} \cdot (S+D)$$
$$D = 255 - brightness$$

6.1.3.5 Colony morphology

Microbial colonies, such as those of *C. albicans* or *P. aeruginosa* often form structures that start from the center of the colony and end in its outer perimeter. Iris detects colony structure complexity by traversing colony pixels in concentric circles, starting at the colony center. Brightness levels of pixels within such a circle feature valleys and peaks, which coincide with colony structures (see **Figure 13**). Since lighting differences can account for brightness peak height, Iris counts the number of peaks in a binary way if they are above a certain brightness threshold.

6.2 Acquiring quantitative phenotypes of Gram-negative envelope integrity

With respect to the methods used in analyzing agar plates images for chromophore detection, I direct the reader to the Materials and Methods section of our published manuscript (Paradis-Bleau et al., 2014). I then built upon this approach and optimized the envelope integrity screen for quantitative measurements over prolonged timespans. The current paragraph thus details only the Materials and Methods used in the part that is yet to be published.

6.2.1 Bacterial strains and plasmids used

In this study I used an ordered *E. coli* mutant library (Nichols et al., 2011) that includes the KEIO collection (Baba et al., 2006), as well as a collection of mutants with hypomorphic alleles of essential genes and mutants lacking genes for small RNAs. Briefly, the Keio collection consists of 3985 single gene deletions strains using the *E. coli* K-12 BW25113 as background strain.

The Keio collection background strain (BW25113) is lacking the *lacZ* gene, which encodes for the β-galactosidase enzyme. Since the permeability assay depends on β-galactosidase activity, I used a strain containing a mobile plasmid, encoding the lacZ gene under the lactose promoter (pCB112) (Paradis-Bleau et al., 2014).

6.2.2 Mutant library preparation

A copy of the previously described ordered E. coli mutant library (Nichols et al., 2011) stored in 384-well format at -80°C was thawed, pinned onto LB-Kan agar plates using a Singer Rotor robot, and grown overnight at 37°C. After growth overnight, the library was transferred to LB agar plates spread with 100ml of an overnight culture of JA200/pCB112 (donor strain/Plac::*lacZ*, Cam^R). The resulting mating plates were incubated overnight at 37°C, positions corresponding to the ordered library were transferred to LB-Kan-Cam plates, and the plates were incubated again at 37°C overnight. All robotic strain transfers were performed using a Singer Rotor robot (Singer Instruments, Watchet, Somerset, UK).

6.2.3 Agar well plate preparation

LB medium containing 1% NaCl and 2% agar was melted at 55°C and supplemented with 20 μg/ml CPRG, 50μM IPTG, and 20μg/ml chloramphenicol. Subsequent liquid handling steps were performed using a Biomek FXP robot (Beckman Coulter Inc., Brea, CA, USA). In each pipetting step, 96 wells of one 384-well plate are filled with 70μl of melted agar medium. Pipetting tips were pre-warmed before each step by iterated aspiration and expulsion of boiling water.

The robot executes a pouring cycle protocol that repeats these steps until pouring of 4 well plates is complete. Pipetting tips were exchanged after completion of each pouring cycle. After each cycle completion plates were centrifuged at 500rpm for 30 seconds to remove possible bubbles in the agar. Well plates used were Nunclon 384-well flat-bottom translucent plates (Thermo Fisher

Scientific, Waltham, MA, USA). Plates were left for agar to solidify on an even surface at room temperature for 4 hours and were subsequently stored to avoid evaporation and inoculated 24 hours later.

Although the liquid handling robot is programmed to deposit the same agar volume in each well, few wells were filled with agar volume, principally owing to the viscosity of the melted agar medium. In case a well has less agar volume, its level will inevitably be lower than that of the rest of the plate. This is a problem for the subsequent inoculation step, in which the pinning robot uses a flat pinning pad to inoculate all wells en masse. In this step, all wells will be inoculated, save for those with lower agar level.

After several rounds of optimization of both plate preparation and inoculation, performed by myself and Dr. Manuel Banzhaf (Typas Lab), the rate of non-inoculated wells at the point of writing was on average 2% per plate. As a positive note, these 2% non-inoculated wells no not show a spatial pattern or preference. Together with the use of 4 replicate colonies per mutant, allowed me to acquire all 4 replicates for most mutants, and 3 replicates for about 5% of the mutants.

6.2.4 Quantitative genome-wide membrane permeability screening process

A plasmid containing the lacZ gene was introduced to all mutants in the Keio collection (see 6.2.2 Mutant library preparation). Subsequently, mutants were transferred to agar well plates containing chloramphenicol, CPRG and IPTG using a Singer Rotor (Singer Instruments, Watchet, Somerset, UK). Each of 12 KEIO library plates (1 clone per mutant) was copied to 4 replicate agar-filled well plates to be used for screening using long-pin pads. Singer Rotor settings were set to 55% pressure to accommodate for the surface tension in each well so that the agar was not pierced.

Bacterial colony growth, as well as CPRG turnover was measured over time using a Tecan Infinite M1000 plate reader (Tecan Group Ltd., Männedorf, Zürich, Switzerland). Immediately after inoculation, plates were monitored hourly for 60 hours at room temperature.

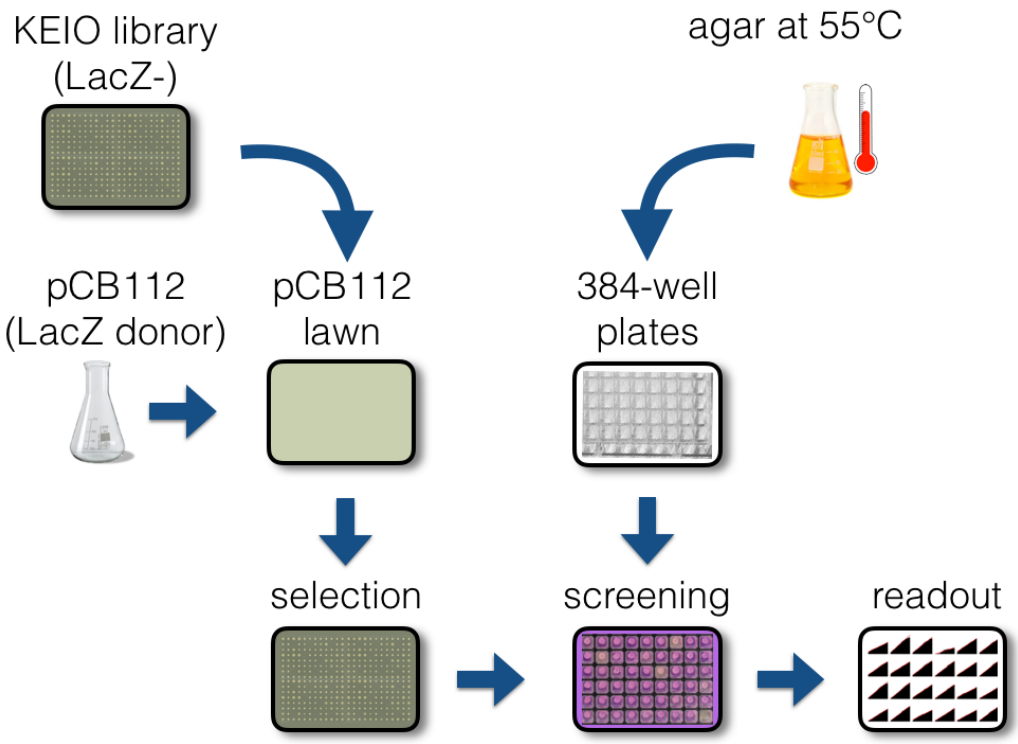

Figure 45: Ordered mutant library screening for membrane defects, overview of the screening procedure.

6.2.5 Readout and data analysis software

Data acquisition was performed using the Tecan i-control software version 1.10.4.0 (Tecan Group Ltd., Männedorf, Zürich, Switzerland). After data acquisition, all data were processed with in-house software written using the statistical analysis platform R (Team, 2014). Data processing steps include the import of data from tables generated by the Tecan i-control software, identification of empty wells, and calculation of the permeability readout. Subsequent analysis and data interpretation was also performed using the same in-house software tools and scripts.

6.3 Exploring domain-protein coevolution among cell wall related proteins

Two of the bifunctional PBPs of *E. coli* (PBP1A and PBP1B) feature domains that serve as docking regions for cognate lipoproteins that activate PBP enzymatic function upon binding to their respective domains. Following the observation that these domains, and their cognate lipoproteins co-occur across species, I set forth to explore the domain content plasticity in core process proteins, with the aim of uncovering their domains' interacting partners.

6.3.1 Core process proteins explored in this study

In the present study I limited the scope of co-evolution analysis to domains in modular proteins belonging to the core process of cell wall biosynthesis and remodeling. I explored in total the domain content, as well as their possible interacting partners of more than 50 modular proteins across 20 different species. Proteins explored within the scope of this work are all proteins containing known cell wall-related enzymatic domains and can be found in Table 1 (see page 72).

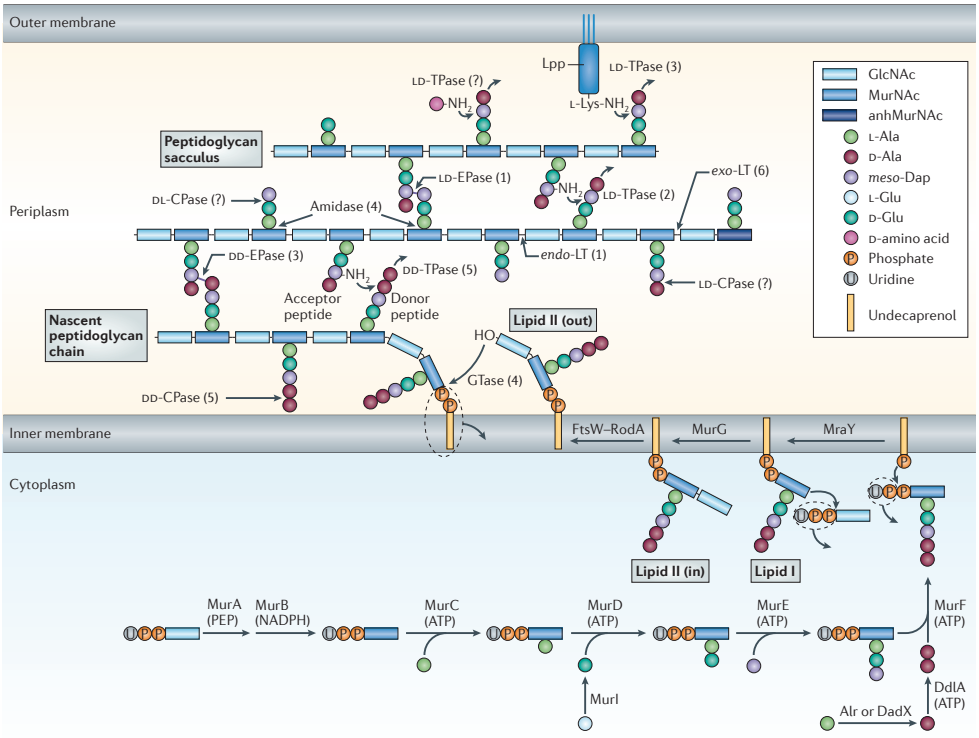

Figure 46: Proteins used in this study are involved in biosynthesis, attachment and degradation of peptidoglycan cell wall. Figure is reused from (Typas et al., 2012), copyright holder Nature Publishing Group, used with copyright holder permission.

I acquired this set of proteins and homologues across species, by analyzing the set of all cell wall-related proteins in *E. coli*. After acquiring their catalytic domain content, I used each of the cell wall-related catalytic domains to search for proteins featuring those domains in 20 bacterial species across the bacterial kingdom. These species were selected for species with complete sequence, as well as their use as model organisms and are presented in Table 2 (see page 73). Subsequently, for each of these retrieved proteins across species, I analyzed their domain content (see 6.3.4) to acquire:

- Domains that are annotated as catalytically active
- Domains of unknown function (DUF)
- Domains of less reliability (PfamB)
- Protein regions that are not annotated as domains

6.3.2 Positive control domain-protein pairs

In the current study I use 3 domain-cofactor protein pairs as positive controls to optimize the computational pipeline:

- LpoA – ODD domain in PBP1A
- LpoB – UB2H domain in PBP1B
- YfhM – BiPBP_C domain in PBPC

Lipoproteins, LpoA and LpoB were shown to bind to and activate their cognate Penicillin Binding Proteins, PBP1A and PBP1B respectively (Egan et al., 2014; Typas et al., 2010). Thus binding was specific to two non-catalytic domains of PBP1A and PBP1B, named ODD and UB2H correspondingly.

An earlier study found that the *yfhM*-like bacterial α2-macroglobulin genes feature strikingly similar phylogenetic distributions with *pbpC* (Budd et al., 2004). PbpC encodes for a PBP that features both catalytic domains, but only its transglycosylase activity has been shown (Schiffer and Holtje, 1999). Also its transpeptidase domain is believed to be non-functional, since it does not show high affinity to most beta-lactams (Schiffer and Holtje, 1999). YfhM on the other hand has been shown to protect the bacterial cell against host peptidases (Doan and Gettins, 2008; Garcia-Ferrer et al., 2015).

To further strengthen the notion that *yfhM* is functionally linked with *pbpC*, it was shown that the two genes are usually in the same operon, or co-transcribed

from the same promoter. Interestingly, in contrast to those of the Lpo pairs, the phylogenetic distribution of *yhfM-pbpC* is not contiguous clades of the bacterial tree, but follows a patchy distribution, suggesting that these genes are acquired through horizontal gene transfer. PbpC typically features a non-catalytic domain, termed BiPBP_C, which is predicted by a secondary structural prediction algorithm PROF to be entirely in beta-fold (Rost, 2001). Although there is no biochemical evidence directly linking proteins YfhM and PbpC, the genetic data make a compelling argument for a possible functional link.

There are several more pairs of cell wall-related proteins that possibly interact, with mounting evidence of such interactions. Examples include the interaction of protein EnvC, an activator of PG hydrolases AmiA and AmiB (Uehara et al., 2010). Another PG hydrolase, AmiC involved in septal splicing, was shown to be regulated by lipoprotein NlpD (Uehara et al., 2010). More recently, protein CpoB was shown to bind both to PBP1B and the Tol system, thus linking the functions of division-related PG assembly to OM constriction (Gray et al., 2015).

6.3.3 Protein phylogenetic distributions

In this study, I used the EggNOG database to acquire protein presence/absence data across species. The EggNOG database consists of groups of orthologous proteins, grouped in an unsupervised manner by means of reciprocally high degree of sequence similarity. The latest iteration of the EggNOG database (v4.5) (Huerta-Cepas et al., 2016) uses genomic sequence information from more than 3600 species. Sequences for almost 15 million proteins across all these species are pairwise aligned using the FASTA algorithm, and their sequence similarity is used by a clustering algorithm to group proteins into clusters of orthologous genes (COGs).

The COGs in EggNOG database are organized in different taxonomic levels, for instance gproNOG are gene clusters that were calculated by using proteins of species only in the Gamma-proteobacteria class. The latest version of EggNOG (v4.5) is organized in more than 100 taxonomic levels (107 in version 4.0), including a level that includes all species used (COG level). In this study I started from using the COG level but expanded to using all taxonomic levels, as it became clear that orthologous gene clusters are often automatically split into smaller subgroups, or aggregated into larger subgroups that do not reflect the

phylogenetic distribution of some of the controls. By using specific taxonomic levels, I ensure that this splitting or grouping was done with respect to the diversity within the particular taxonomic level, and thus inaccurate dividing or aggregating groups of genes is reduced.

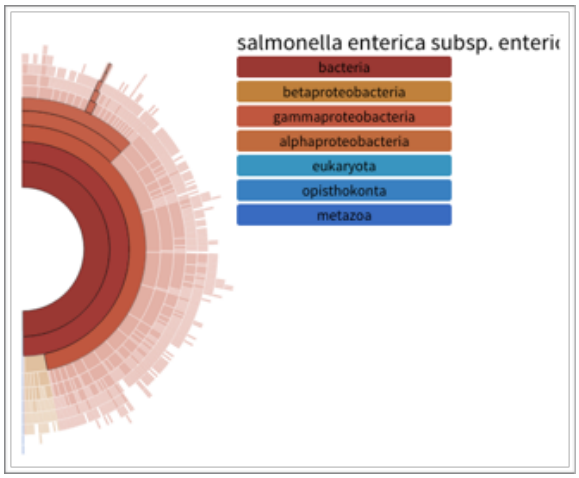

Figure 47: Visualization of LpoA COG phylogenetic distribution, highlighted species is Salmonella enterica subspecies enterica, serovar Typhimurium

6.3.4 Modular protein domain annotations and inter-domain regions

As a starting point in exploring protein modularity and domain content plasticity, I used the Pfam domain annotation database (Finn et al., 2014). The Pfam database is a large collection of protein domains, each calculated using multiple sequence alignments and represented using hidden Markov models (HMMs) (Eddy, 1998). The HMMs of each protein domain are then further used to detect domains in protein sequences, using HMM search algorithms (Eddy, 2009; Eddy, 2011)}. Moreover, the Pfam database holds the sequences of each protein, annotated with their domain content, which is readily available both through their web interface, as well as a programmatic interface.

In this study, I used the programmatic access in Pfam database, in order to acquire large numbers of protein domain annotations. These domain annotations are essential to this analysis, since I can process them to provide the following 3 key pieces of information:

- The domain content plasticity of a protein family of interest,

- The phylogenetic distribution of a domain of interest within a protein family
- The domain genomic positions for each protein in a family, which is used to acquire inter-domain (i.e. non-domain-annotated) regions

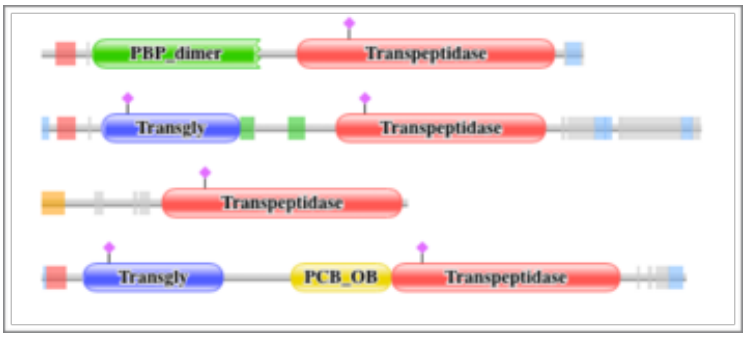

Figure 48: Example of domain architecture visualization in Pfam: domain architectures containing the Transpeptidase domain. Source: Pfam (Finn et al., 2014)

The Pfam database is programmatically accessible through a REST interface. This interface was also used in the current study to acquire protein domain annotations en masse. More information on the Pfam programmatic interface can be found on their website, under http://pfam.xfam.org/help.

Finally, the Pfam database holds information also on protein regions that were found to be conserved across species, and annotated as a domain using an alternative algorithm, called ADDA (Heger et al., 2005). These regions, called PfamB domains are generally considered of lower quality. However, they serve as a useful indicator of a conserved region. PfamB domains are included in this study, and inter-domain regions typically exclude PfamB domains. It's worth noting that PfamB domains are no longer supported in the Pfam database, since version 28.0 (released June 2015).

The current work was performed using Pfam version 27.0, which still features PfamB domains. Newer versions of the Pfam database however include domain definitions for PBP domains, including the ODD domain, which can be found under the name PCB_OB. Such domains were often annotated as DNA binding domains, however after personal communication with the Pfam authors, Alex Bateman and Rob Finn these domains were prioritized to be added in the Pfam database as separate entries. Finally, the addition of this and other PBP domains to the Pfam database was performed in collaboration with Ruth Eberhard, a Pfam curator.

6.3.5 Domain phylogenetic distributions

The Pfam database holds Hidden Markov Models (HMM) for each of the PfamA domain families. In order to retrieve domain phylogenetic distributions for a domain of interest in the EggNOG v4 sequence database, I first programmatically acquired the pre-calculated HMM for this domain family from the Pfam database (Finn et al., 2014). Next, I used this HMM as input to the *hmmsearch* algorithm version 3.1b1 and ran against the EggNOG v4 sequence database. To recapitulate the results presented in the Pfam database, I ran the hmmsearch algorithm using the same parameters as defined in the Pfam website.

6.3.6 Phylogenetic distributions for PfamB domains and inter-domain regions

In order to acquire phylogenetic distributions for PfamB domains and inter-domain regions (regions with no domain annotation), I followed a similar approach to this of retrieving phylogenetic distributions for annotated domains. The difference lies with the fact that there is no pre-calculated HMM in the Pfam database for PfamB domains, and inter-domain regions.

For inter-domain regions, I used the *jackHMMER* algorithm version 3.1b1 (Johnson et al., 2010) with a strict E-value threshold (10^{-12}) to directly search on the EggNOG v4 sequence database. JackHMMER iteratively searches against a sequence database, building an HMM from a query sequence by using closely related homologue sequences. For each jackHMMER step this HMM is recalculated by using the retrieved sequences. This strict jackHMMER threshold was used in order to avoid the HMM profile iteratively diverging from the original query sequence; this is referred to in the JackHMMER documentation as an *iterative walk in sequence space*.

For the PfamB domains I used the multiple sequence alignment provided by the Pfam database to calculate an HMM. To this end, I used the *hmmbuild* algorithm version 3.1b1 (Eddy, 2009; Eddy, 2011). The provided HMM was then used to search against the EggNOG v4 sequence database as described above.

6.3.7 Physical interaction and genomic context data

To evaluate the results of the present pipeline, and to make the results presentation more user friendly, I also incorporated any known information that

may corroborate our results. Specifically, I include information on known physical interactions, or genomic context between the protein holding the domain of interest and every possible interaction partner. To do this I retrieve all physical interaction, text mining, as well as genomic context data from the 10th iteration of the STRING database (Szklarczyk et al., 2014).

6.3.8 Protein membrane localization

In order to aid the user I annotate ranked gene clusters in the results as to whether they include proteins annotated in the Uniprot database (EMBL et al., 2013) to have a signal sequence. Signal peptide-featuring proteins are targeted to the endoplasmic reticulum in eukaryotic cells. In bacteria, signal peptide proteins are generally targeted to the cell envelope, or tagged for secretion (Masi and Wandersman, 2010).

Since the proteins targeted in this study are all present in the periplasmic space, I expect that their potential interacting partners are also localized in the periplasm. Thus presenting signal peptide information to the user adds a useful localization component to the results. The Uniprot database automatically annotates proteins holding a signal sequence by making use of 4 existing algorithms:

- Phobius (Käll et al., 2007)
- Predotar (Small et al., 2004)
- SignalP (Bendtsen et al., 2004; Emanuelsson et al., 2007)
- TargetP (Emanuelsson et al., 2007; Emanuelsson et al., 2000)

If any of these 2 algorithms return a positive result, then the protein in Uniprot is annotated as a signal peptide.

6.3.9 Gene Ontology enrichment analysis

The Gene Ontology enrichment analysis in this study was performed using an in-house developed function in R. This function uses Fisher's exact test (Fisher, 1922) to calculate the enrichment in occurrences of every single GO term in the foreground compared to the background set. Results are corrected for multiple testing by means of the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995), implemented by the R package *frdtool* (Strimmer, 2008)

6.3.10 Co-occurrence across species

By acquiring large-scale information of protein family phylogenetic distribution, one can then compare this information to domain phylogenetic distribution. Keeping in mind the positive control paradigm (PBP domain and cognate lipoprotein) I use metrics that quantify the degree in which a particular protein presence across species overlaps with the domain presence in a protein family of interest. Co-occurrence methods, such as mutual information (MI), have been also used by the STRING database in quantifying the degree of protein-level co-occurrence across species (von Mering et al., 2003).

For the remainder of this section, I will use the term domain referring always to a domain in a modular protein of interest. Furthermore, I will refer to the set of species that feature protein or domain in their genome as the phylogenetic distribution of this protein or domain.

Since I am searching for a protein that matches the phylogenetic distribution of a domain of interest, I can easily transform the question problem into an Information Retrieval problem, whereby the domain phylogenetic distribution is equivalent to the ground truth set of species and each protein family's phylogenetic distribution is equivalent to a prediction of this ground truth set. Domain annotations are retrieved as described from the Pfam database and filtered to keep only domains in a particular family of interest. As protein family definition, I will be using the defined clusters of orthologous genes provided by the EggNOG database (see 6.3.3).

6.3.10.1 F-measure

The metrics that I employ to quantify phylogenetic distribution overlap between a domain and a protein consist of the classic information retrieval measures: Precision, Recall and F-measure (Van Rijsbergen, 1979). The F-measure summarizes both Precision and Recall, and by means of the β factor can be weighed according to the metric that one is interested in the most. For example the F2 measure (β=2), used throughout in this work, weighs recall twice as much as precision. The reason I am penalizing false positives is because species that feature a protein but do not feature the domain were empirically found to be more often in the positive control domain-protein pairs than the opposite case. The underlying reason was attributed to detection accuracy of protein, and protein domain content across species. As mentioned above, I found that the EggNOG

database often contains imperfectly defined gene groups. For the positive control proteins it was often the case that few species were erroneously included in their EggNOG gene group. Thus penalizing these false positive species in COGs more than false negatives resulted in a better ranking of the positive control domain-protein pairs.

6.3.10.2 Mutual Information

In order to measure domain and protein co-occurrence I also employ another information theory measure, the point-wise Mutual Information. Point-wise Mutual Information is already used in calculating co-occurrence in STRING. Mutual information is also widely used -among other fields- in Natural Language Processing (NLP); example applications include synonym discovery (Manning et al., 1999), and opinion extraction from product reviews (Popescu and Etzioni, 2007). Arguably, NLP applications have many equivalent applications to tasks in bioinformatics. An example includes word co-occurrence across bodies of text, of which the equivalent in many aspects is protein co-occurrence across species.

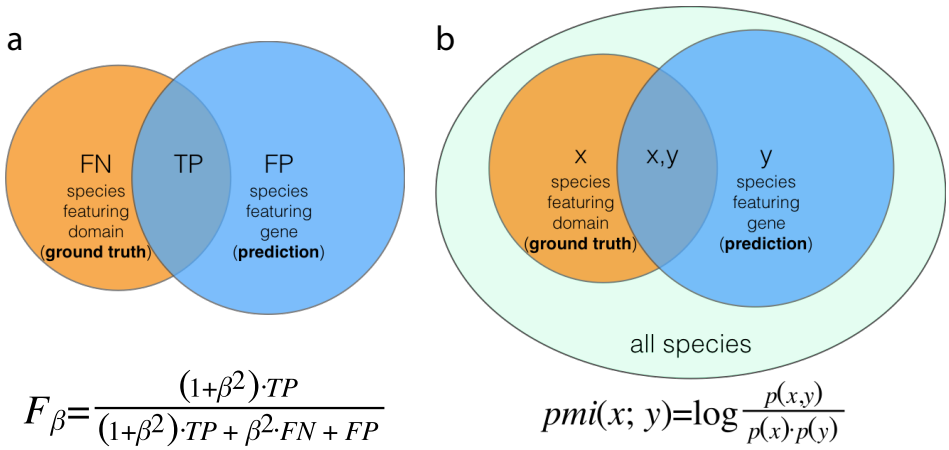

Figure 49: Visual representation of the co-occurrence measures used in this study. Domain phylogenetic distribution is considered to be the ground truth, while gene cluster phylogenetic distributions are assessed by how well they predict this ground truth. TP stands for true positive, FN is false negative, and FP is false positive.

6.3.11 Phylogenetic tree similarity

Co-evolving species pairs, such as pairs under a prey-predator relationship, often feature similar phylogenetic trees. Indeed an often-used example is Charles Darwin's prediction of a moth with a longer proboscis after observing an orchid with a longer spur (Darwin, 1862), a prediction that was verified in 1903. Such

species-level phylogenetic tree similarities can be paralleled at the molecular level to similarities at the phylogenetic trees of proteins or -in this case- protein domains.

6.3.11.1 MirrorTree

MirrorTree was the first method developed by Florencio Pazos and Alfonso Valencia in 2001 (Pazos and Valencia, 2001). The algorithm makes use of pre-calculated pairwise distances within a protein family of interest, which is typically derived through a multiple sequence alignment (MSA) algorithm, such as ClustalW (Larkin et al., 2007; Thompson et al., 1994). Interestingly, the algorithm operates on these pairwise distances without the need for tree calculation, thus making this (and following) "tree comparison" methods misnomers.

In the case that more than one proteins of the family of interest exist in a certain species (paralogous sequences), only one needs to be selected. As criteria for paralog selection, the authors selected the sequence which was closer to the master sequence for the protein family, found in the HSSP database (Dodge et al., 1998). After removal of paralogous sequences, pairwise sequence distances refer to sequence distances across pairs of species.

In the last stage of the algorithm, within protein family pairwise distances are compared to those of other protein families. In order to compare with the set of distances across species to those of another protein family, the algorithm selects only the species pairs that exist in both families. Importantly, a limited overlap between species pairs is observed to artificially inflate the algorithm results. The method describes setting an overlap threshold (typically dozens of species) below which the selected protein families cannot be compared with this method. Finally, the algorithm compares the within-family distances of two protein families by means of a linear correlation coefficient. The reported correlation coefficient value varies between -1 and 1, with 1 denoting identical inter-species distances among the compared proteins.

A drawback of the MirrorTree approach is that a random pair or protein families is expected to show a high degree of phylogenetic tree similarity. This is explained by the fact that the phylogenetic trees (or protein pairwise distances) are heavily influenced by speciation events. In other words, two protein sequence distance-derived trees will tend to be similar to each other, because they are both similar to the underlying Tree of Life (TOL). Refinements on the original

MirrorTree method, such as the TOL MirrorTree (Pazos and Valencia, 2001), and ContextMirror (Pazos et al., 2008) largely had to do with mitigating this issue.

6.3.11.2 TOL-MirrorTree

TOL-MirrorTree was introduced by Pazos and Valencia in 2005 as an improvement on the original MirrorTree method (Pazos and Valencia, 2001). MirrorTree's inherent issue is that it can detect spurious tree similarities owing to the fact that any protein-distance derived tree is influenced by the underlying speciation events. TOL-MirrorTree attempts to solve this issue by subtracting the effect of these speciation events on the protein-distance derived phylogenetic trees.

TOL-MirrorTree follows an identical procedure to the original MirrorTree (see 6.3.11.1), adding a subtraction step independently performed in each protein tree before the trees are compared. The authors estimate the sequence drift by calculating the 16S rRNA inter-species distances. The derived 16S rRNA tree (referred to as the Tree of Life - TOL) serves as a proxy to the contribution of speciation events in protein distances across species. Finally, when applied to the same dataset, the TOL-MirrorTree was shown to outperform the original MirrorTree method.

The original TOL-MirrorTree uses the 16S rRNA distances to compute the TOL effect. However advances in prokaryotic lineage determination allow us to use a set of 40 single-copy marker genes (Mende et al., 2013) in order to better approximate the sequence drift due to speciation events.

6.3.11.3 Context Mirror

ContextMirror builds on top of MirrorTree's output, and defines the co-evolution profile of a protein family as a vector of MirrorTree results across all other proteins families. Subsequently, the similarity between pairs of co-evolution profiles is calculated by means of their Pearson's correlation coefficient. Finally, ContextMirror evaluates the influence of third proteins on the coevolution of a given pair of proteins, by calculating the partial correlation coefficients of each such pair given every third protein.

While in the original MirrorTree, pairwise protein distances were calculated directly from the Multiple Sequence Alignment algorithm ClustalW (Larkin et al., 2007; Thompson et al., 1994), ContextMirror requires that these distances are further clustered into a phylogenetic tree using the Neighbor-Joining algorithm

(Saitou and Nei, 1987). Subsequently, this tree is decomposed back into pairwise protein distances, which are calculated by summing the branch lengths separating each pair of proteins. MirrorTree is then ran using these distances for all pairs of protein families in the analysis, resulting in an N-by-N symmetric matrix, where N is the number protein families, and each value is the result of their MirrorTree comparison. As in the original MirrorTree, a minimum set of species is required to produce a result; authors used a threshold of 15 species, which is what I also use in my implementation.

In the following step ContextMirror treats each row- or column-wise vector in the above matrix as the coevolution signature of the protein family corresponding to this vector. All such pairwise coevolution signature vectors are then compared by means of a simple Pearson's correlation coefficient, with correlations over a p-value threshold (P-value<10^{-5}) stored in a symmetric signature correlation matrix. The rationale behind the correlation of coevolution signatures lies with the fact that each MirrorTree value is likely high. However, the similarity of coevolution signatures calculated across all proteins is much less affected by spurious MirrorTree results. Thus in this elegant step ContextMirror overcomes MirrorTree's shortcomings with respect to the influence of speciation events.

In the final step, ContextMirror assesses to what extent the co-evolution of a pair of proteins is specific or shared among a group of proteins. This is performed on top of the previous signature correlation matrix, by calculating the correlation of profiles of all protein pairs AB given every third protein. This final step is useful in disentangling pairwise protein co-evolution effects from those mediated for instance due to protein complex membership.

In the current work, I am using Context Mirror to acquire protein-domain correlation signatures that overcome the limitations of the original MirrorTree. Thus the final step of ContextMirror pertaining more to protein complex discovery was neither implemented nor used in the current work, however I encourage the interested reader to refer to the original publication (Pazos et al., 2008).

Bibliography

Albataineh, M.T., Lazzell, A., Lopez-Ribot, J.L., and Kadosh, D. (2014). Ppg1, a PP2A-type protein phosphatase, controls filament extension and virulence in Candida albicans. Eukaryot Cell *13*, 1538-1547.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. Journal of molecular biology *215*, 403-410.

Anantharaman, V., and Aravind, L. (2003). Evolutionary history, structural features and biochemical diversity of the NlpC/P60 superfamily of enzymes. Genome biology *4*, R11.

Apic, G., Gough, J., and Teichmann, S.A. (2001). Domain combinations in archaeal, eubacterial and eukaryotic proteomes. Journal of Molecular Biology *310*, 311-325.

Arumugam, M., Raes, J., Pelletier, E., Le Paslier, D., Yamada, T., Mende, D.R., Fernandes, G.R., Tap, J., Bruls, T., Batto, J.-M., *et al.* (2011). Enterotypes of the human gut microbiome. Nature *473*, 174-180.

Baba, T., Ara, T., Hasegawa, M., Takai, Y., Okumura, Y., Baba, M., Datsenko, K.K.A., Tomita, M., Wanner, B.L.B.L., Mori, H., *et al.* (2006). Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. Mol Syst Biol *2*, 2006 0008.

Banzhaf, M., van den Berg van Saparoea, B., Terrak, M., Fraipont, C., Egan, A., Philippe, J., Zapun, A., Breukink, E., Nguyen-Distèche, M., den Blaauwen, T., *et al.* (2012). Cooperativity of peptidoglycan synthases active in bacterial cell elongation. Molecular Microbiology *85*, 179-194.

Barnhart, M.M., and Chapman, M.R. (2006). Curli biogenesis and function. Annu Rev Microbiol *60*, 131-147.

Barreteau, H., Magnet, S., El Ghachi, M., Touze, T., Arthur, M., Mengin-Lecreulx, D., and Blanot, D. (2009). Quantitative high-performance liquid chromatography analysis of the pool levels of undecaprenyl phosphate and its derivatives in bacterial membranes. J Chromatogr B Analyt Technol Biomed Life Sci *877*, 213-220.

Bastidas, R.J., and Heitman, J. (2009). Trimorphic stepping stones pave the way to fungal virulence. Proceedings of the National Academy of Sciences *106*, 351-352.

Behrens, S., Maier, R., de Cock, H., Schmid, F.X., and Gross, C.A. (2001). The SurA periplasmic PPIase lacking its parvulin domains functions in vivo and has chaperone activity. EMBO J *20*, 285-294.

Bendtsen, J.D., Nielsen, H., Von Heijne, G., and Brunak, S. (2004). Improved prediction of signal peptides: SignalP 3.0. Journal of Molecular Biology *340*, 783-795.

Benjamini, Y., and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. In Journal of the Royal Statistical Society Series B (Methodological), pp. 289 - 300.

Benjamini, Y., and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under depencency. The Annals of Statistics *29*, 1165-1188.

Bernstein, H.D. (2011). The double life of a bacterial lipoprotein. Molecular microbiology *79*, 1128-1131.

Bertsche, U., Kast, T., Wolf, B., Fraipont, C., Aarsman, M.E., Kannenberg, K., von Rechenberg, M., Nguyen-Disteche, M., den Blaauwen, T., Höltje, J.-V., *et al.* (2006). Interaction between two murein (peptidoglycan) synthases, PBP3 and PBP1B, in Escherichia coli. Mol Microbiol *61*, 675-690.

Bos, M.P., Robert, V., and Tommassen, J. (2007). Biogenesis of the Gram-Negative Bacterial Outer Membrane. Annu Rev Microbiol *61*, 191-214.

Braun, V., and Wolff, H. (1970). The murein-lipoprotein linkage in the cell wall of Escherichia coli. European journal of biochemistry / FEBS *14*, 387-391.

Brochado, A.R., and Typas, A. (2013). High-throughput approaches to understanding gene function and mapping network architecture in bacteria. Current opinion in microbiology *16*, 199-206.

Budd, A., Blandin, S., Levashina, E.A., and Gibson, T.J. (2004). Bacterial alpha2-macroglobulins: colonization factors acquired by horizontal gene transfer from the metazoan genome? Genome biology *5*, R38.

Cabeen, M.T., Leiman, S.A., and Losick, R. (2016). Colony-morphology screening uncovers a role for the P seudomonas aeruginosa nitrogen-related phosphotransferase system in biofilm formation. Molecular Microbiology *99*, 557-570.

Cho, S.H., Szewczyk, J., Pesavento, C., Zietek, M., Banzhaf, M., Roszczenko, P., Asmar, A., Laloux, G., Hov, A.K., Leverrier, P., *et al.* (2014). Detecting envelope stress by monitoring beta-barrel assembly. Cell *159*, 1652-1664.

Cleveland, W.S. (1981). LOWESS: A program for smoothing scatterplots by robust locally weighted regression. American Statistician *35*, 54.

Conrad, C., and Gerlich, D.W. (2010). Automated microscopy for high-content RNAi screening. The Journal of Cell Biology *188*, 453-461.

Costanzo, M., Baryshnikova, A., Bellay, J., Kim, Y., Spear, E.D., Sevier, C.S., Ding, H., Koh, J.L., Toufighi, K., Mostafavi, S., *et al.* (2010). The genetic landscape of a cell. Science *327*, 425-431.

D'Andrea, L.D., and Regan, L. (2003). TPR proteins: The versatile helix. In Trends in Biochemical Sciences, pp. 655-662.

D'Argenio, D.A., Calfee, M.W., Rainey, P.B., and Pesci, E.C. (2002). Autolysis and autoaggregation in Pseudomonas aeruginosa colony morphology mutants. Journal of bacteriology *184*, 6481-6489.

Darwin, C. (1862). On the contrivances by which British and foreign orchids are fertilized by insects, and on the good effects of intercrossing. John Murray, 365.

Davies, J., and Davies, D. (2010). Origins and Evolution of Antibiotic Resistance. Microbiology and Molecular Biology Reviews *74*, 417-433.

Davis, D., Edwards, J.E., Jr., Mitchell, A.P., and Ibrahim, A.S. (2000). Candida albicans RIM101 pH response pathway is required for host-pathogen interactions. Infect Immun *68*, 5953-5959.

de Vries, S.J., van Dijk, M., and Bonvin, A.M.J.J. (2010). The HADDOCK web server for data-driven biomolecular docking. Nature protocols *5*, 883-897.

Delcour, A.H. (2009). Outer Membrane Permeability and Antibiotic Resistance. Biochim Biophys Acta *1794*, 808-816.

Deutschbauer, A., Price, M.N., Wetmore, K.M., Shao, W., Baumohl, J.K., Xu, Z., Nguyen, M., Tamse, R., Davis, R.W., and Arkin, A.P. (2011). Evidence-based annotation of gene function in Shewanella oneidensis MR-1 using genome-wide fitness profiling across 121 conditions. PLoS Genet *7*, e1002385.

Dietrich, L.E.P., Price-Whelan, A., Petersen, A., Whiteley, M., and Newman, D.K. (2006). The phenazine pyocyanin is a terminal signalling factor in the quorum sensing network of Pseudomonas aeruginosa. Molecular microbiology *61*, 1308-1321.

Doan, N., and Gettins, P.G.W. (2008). alpha-Macroglobulins are present in some gram-negative bacteria: characterization of the alpha2-macroglobulin from Escherichia coli. The Journal of biological chemistry *283*, 28747-28756.

Dodge, C., Schneider, R., and Sander, C. (1998). The HSSP database of protein structure-sequence alignments and family profiles. Nucleic Acids Research *26*, 313-315.

Donati, C., Hiller, N.L., Tettelin, H., Muzzi, A., Croucher, N.J., Angiuoli, S.V., Oggioni, M., Dunning Hotopp, J.C., Hu, F.Z., Riley, D.R., *et al.* (2010). Structure and dynamics of the pan-genome of Streptococcus pneumoniae and closely related species. Genome Biology *11*, R107.

Donlan, R.M., and Costerton, J.W. (2002). Biofilms: Survival Mechanisms of Clinically Relevant Microorganisms. Clinical Microbiology Reviews *15*, 167-193.

Eddy, S. (1998). Profile hidden Markov models. Bioinformatics *14*, 755-763.

Eddy, S.R. (2009). A new generation of homology search tools based on probabilistic inference. Genome informatics International Conference on Genome Informatics *23*, 205-211.

Eddy, S.R. (2011). Accelerated profile HMM searches. PLoS computational biology *7*, e1002195.

Egan, A.J.F., Biboy, J., van't Veer, I., Breukink, E., and Vollmer, W. (2015). Activities and regulation of peptidoglycan synthases. Philosophical transactions of the Royal Society of London Series B, Biological sciences *370*, 193-232.

Egan, A.J.F., Jean, N.L., Koumoutsi, A., Bougault, C.M., Biboy, J., Sassine, J., Solovyova, A.S., Breukink, E., Typas, A., Vollmer, W., *et al.* (2014). Outer-membrane lipoprotein LpoB spans the periplasm to stimulate the peptidoglycan synthase PBP1B. Proceedings of the National Academy of Sciences of the United States of America *111*, 8197-8202.

El Feghaly, R.E., Bangar, H., and Haslam, D.B. (2015). The molecular basis of Clostridium difficile disease and host response. Current opinion in gastroenterology *31*, 24-29.

Emanuelsson, O., Brunak, S., von Heijne, G., and Nielsen, H. (2007). Locating proteins in the cell using TargetP, SignalP and related tools. Nature protocols *2*, 953-971.

Emanuelsson, O., Nielsen, H., Brunak, S., and von Heijne, G. (2000). Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. Journal of molecular biology *300*, 1005-1016.

EMBL, Bioinformatics, S.S.I.o., and (PIR), P.I.R. (2013). UniProt. In Nucleic acids research, pp. 41: D43-D47.

Eustice, D.C., Feldman, P.A., Colberg-Poley, A.M., Buckery, R.M., and Neubauer, R.H. (1991). A sensitive method for the detection of beta-galactosidase in transfected mammalian cells. BioTechniques *11*, 739-740,742-743.

Evans, Margery L., Chorell, E., Taylor, Jonathan D., Åden, J., Götheson, A., Li, F., Koch, M., Sefer, L., Matthews, Steve J., Wittung-Stafshede, P., *et al.* (2015). The Bacterial Curli System Possesses a Potent and Selective Inhibitor of Amyloid Formation. Molecular Cell *57*, 445-455.

Fernandez, L., and Hancock, R.E.W. (2012). Adaptive and Mutational Resistance: Role of Porins and Efflux Pumps in Drug Resistance. Clinical Microbiology Reviews *25*, 661-681.

Finn, R.D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R.Y., Eddy, S.R., Heger, A., Hetherington, K., Holm, L., Mistry, J., *et al.* (2014). Pfam: The protein families database. In Nucleic Acids Research.

Firczuk, M., Bochtler, M., Anantharaman, V., Aravind, L., Baker, J., Liu, C., Dong, S., Pritchard, D., Bateman, A., Rawlings, N., *et al.* (2007). Folds and activities of peptidoglycan amidases. FEMS microbiology reviews *31*, 676-691.

Fisher, R. (1922). On the interpretation of χ2 from contingency tables, and the calculation of P. Journal of the Royal Statistical Society *85*, 87-94.

Fox, E.P., Bui, C.K., Nett, J.E., Hartooni, N., Mui, M.C., Andes, D.R., Nobile, C.J., and Johnson, A.D. (2015). An expanded regulatory network temporally controls Candida albicans biofilm formation. Mol Microbiol *96*, 1226-1239.

Franke, K., Nguyen, M., Hartl, A., Dahse, H.M., Vogl, G., Wurzner, R., Zipfel, P.F., Kunkel, W., and Eck, R. (2006). The vesicle transport protein Vac1p is required for virulence of Candida albicans. Microbiology *152*, 3111-3121.

Frost, A., Elgort, M.G., Brandman, O., Ives, C., Collins, S.R., Miller-Vedam, L., Weibezahn, J., Hein, M.Y., Poser, I., Mann, M., *et al.* (2012). Functional repurposing revealed by comparing S. pombe and S. cerevisiae genetic interactions. Cell *149*, 1339-1352.

Garcia-Ferrer, I., Arêde, P., Gómez-Blanco, J., Luque, D., Duquerroy, S., Castón, J.R., Goulas, T., and Gomis-Rüth, F.X. (2015). Structural and functional insights into Escherichia coli α ₂ -macroglobulin endopeptidase snap-trap inhibition. Proceedings of the National Academy of Sciences *112*, 8290-8295.

Ghafoor, A., Hay, I.D., and Rehm, B.H.A. (2011). Role of exopolysaccharides in Pseudomonas aeruginosa biofilm formation and architecture. Applied and Environmental Microbiology *77*, 5238-5246.

Goebl, M., and Yanagida, M. (1991). The TPR snap helix: a novel protein repeat motif from mitosis to transcription. Trends in Biochemical Sciences *16*, 173-177.

Goffin, C., and Ghuysen, J.M. (1998). Multimodular penicillin-binding proteins: an enigmatic family of orthologs and paralogs. Microbiology and molecular biology reviews : MMBR *62*, 1079-1093.

Gray, A.N., Egan, A.J.F., Van't Veer, I.L., Verheul, J., Colavin, A., Koumoutsi, A., Biboy, J., Altelaar, A.F.M., Damen, M.J., Huang, K.C., *et al.* (2015). Coordination of peptidoglycan synthesis and outer membrane constriction during Escherichia coli cell division. eLife *4*, 1-29.

Guvener, Z.T., and Harwood, C.S. (2007). Subcellular location characteristics of the Pseudomonas aeruginosa GGDEF protein, WspR, indicate that it produces cyclic-di-GMP in response to growth on surfaces. Mol Microbiol *66*, 1459-1473.

Hagan, C.L., Silhavy, T.J., and Kahne, D. (2011). β-Barrel Membrane Protein Assembly by the Bam Complex. Annual Review of Biochemistry *80*, 189-210.

Haiser, H.J., Yousef, M.R., and Elliot, M.A. (2009). Cell Wall Hydrolases Affect Germination, Vegetative Growth, and Sporulation in Streptomyces coelicolor. J Bacteriol *191*.

Hall-Stoodley, L., Costerton, J.W., and Stoodley, P. (2004). Bacterial biofilms: from the Natural environment to infectious diseases. Nature Reviews Microbiology *2*, 95-108.

Han, L., Zheng, J., Wang, Y., Yang, X., Liu, Y., Sun, C., Cao, B., Zhou, H., Ni, D., Lou, J., *et al.* (2016). Structure of the BAM complex and its implications for biogenesis of outer-membrane proteins. Nature Structural & Molecular Biology *23*, 192-196.

Heath, R.J., and Rock, C.O. (1996). Roles of the FabA and FabZ beta-hydroxyacyl-acyl carrier protein dehydratases in Escherichia coli fatty acid biosynthesis. The Journal of biological chemistry *271*, 27795-27801.

Heger, A., Wilton, C.A., Sivakumar, A., and Holm, L. (2005). ADDA: A domain database with global coverage of the protein universe. Nucleic Acids Research *33*.

Hengge, R. (2009). Principles of c-di-GMP signalling in bacteria. Nature Reviews Microbiology *7*, 263-273.

Hillenmeyer, M.E., Fung, E., Wildenhain, J., Pierce, S.E., Hoon, S., Lee, W., Proctor, M., St Onge, R.P., Tyers, M., Koller, D., *et al.* (2008). The chemical genomic portrait of yeast: uncovering a phenotype for all genes. Science *320*, 362-365.

Höltje, J.-V. (1998). Growth of the stress-bearing and shape-maintaining murein sacculus of Escherichia coli. Microbiol Mol Biol Rev *62*, 181-203.

Homann, O.R., Dea, J., Noble, S.M., and Johnson, A.D. (2009). A phenotypic profile of the Candida albicans regulatory network. PLoS genetics *5*, e1000783.

Howarth, P., and Rüger, S. (2004). Evaluation of Texture Features for Content-Based Image Retrieval. Proceedings of Image and Video Retrieval 2004 *3115*, 326--334.

Hu, P., Janga, S.C., Babu, M., D??az-Mej??a, J.J., Butland, G., Yang, W., Pogoutse, O., Guo, X., Phanse, S., Wong, P., *et al.* (2009). Global functional atlas of Escherichia coli encompassing previously uncharacterized proteins. PLoS Biology *7*, e96.

Huang, B., Whitchurch, C.B., and Mattick, J.S. (2003). FimX, a multidomain protein connecting environmental signals to twitching motility in Pseudomonas aeruginosa. Journal of bacteriology *185*, 7068-7076.

Huerta-Cepas, J., Szklarczyk, D., Forslund, K., Cook, H., Heller, D., Walter, M.C., Rattei, T., Mende, D.R., Sunagawa, S., Kuhn, M., *et al.* (2016). eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. Nucleic acids research *44*, D286-293.

Izaki, K., Matsuhashi, M., and Strominger, J.L. (1966). Glycopeptide transpeptidase and D-alanine carboxypeptidase: penicillin-sensitive enzymatic reactions. Proceedings of the National Academy of Sciences of the United States of America *55*, 656-663.

Jacobs, M.A., Alwood, A., Thaipisuttikul, I., Spencer, D., Haugen, E., Ernst, S., Will, O., Kaul, R., Raymond, C., Levy, R., *et al.* (2003). Comprehensive transposon mutant library of Pseudomonas aeruginosa. Proc Natl Acad Sci U S A *100*, 14339-14344.

Jacobsen, A., Hendriksen, R.S., Aaresturp, F.M., Ussery, D.W., and Friis, C. (2011). The Salmonella enterica pan-genome. Microbial ecology *62*, 487-504.

Jain, R., Behrens, A.-J., Kaever, V., and Kazmierczak, B.I. (2012). Type IV pilus assembly in Pseudomonas aeruginosa over a broad range of cyclic di-GMP concentrations. Journal of bacteriology *194*, 4285-4294.

Jenal, U., and Malone, J. (2006). Mechanisms of Cyclic-di-GMP Signaling in Bacteria. Annu Rev Genet *40*, 385-407.

Johnson, L.S., Eddy, S.R., and Portugaly, E. (2010). Hidden Markov model speed heuristic and iterative HMM search procedure. BMC bioinformatics *11*, 431.

Jones, D.T., Buchan, D.W.A., Cozzetto, D., and Pontil, M. (2012). PSICOV: Precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. Bioinformatics *28*, 184-190.

Juan, D., Pazos, F., and Valencia, A. (2008). High-confidence prediction of global interactomes based on genome-wide coevolutionary networks. Proceedings of the National Academy of Sciences of the United States of America *105*, 934-939.

Kahm, M., Hasenbrink, G., Lichtenberg-frate, H., Ludwig, J., and Kschischo, M. (2010). Grofit: Fitting biological growth curves. Journal of Statistical Software *33*, 1-21.

Käll, L., Krogh, A., and Sonnhammer, E.L.L. (2007). Advantages of combined transmembrane topology and signal peptide prediction-the Phobius web server. Nucleic Acids Research *35*.

Kapitzky, L., Beltrao, P., Berens, T.J., Gassner, N., Zhou, C., Wuster, A., Wu, J., Babu, M.M., Elledge, S.J., Toczyski, D.*, et al.* (2010). Cross-species chemogenomic profiling reveals evolutionarily conserved drug mode of action. Mol Syst Biol *6*, 451.

Khalaf, R.A., and Zitomer, R.S. (2001). The DNA binding protein Rfg1 is a repressor of filamentation in Candida albicans. Genetics *157*, 1503-1512.

Kim, M.J., Kil, M., Jung, J.H., and Kim, J. (2008). Roles of Zinc-responsive transcription factor Csr1 in filamentous growth of the pathogenic Yeast Candida albicans. J Microbiol Biotechnol *18*, 242-247.

Kim, Y., and Mylonakis, E. (2011). Killing of Candida albicans filaments by Salmonella enterica serovar Typhimurium is mediated by sopB effectors, parts of a type III secretion system. Eukaryot Cell *10*, 782-790.

Knuth, D.E. (1997). Art of Computer Programming, Volume 1: Fundamental Algorithms. Journal of the American Statistical Association *1*, 435-455.

Kuhn, H.M., Meier-Dieter, U., and Mayer, H. (1988). ECA, the enterobacterial common antigen. FEMS Microbiol Rev *4*, 195-222.

Lane, S., Zhou, S., Pan, T., Dai, Q., and Liu, H. (2001). The basic helix-loop-helix transcription factor Cph2 regulates hyphal development in Candida albicans partly via TEC1. Molecular and cellular biology *21*, 6418-6428.

Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., Mcgettigan, P.A., McWilliam, H., Valentin, F., Wallace, I.M., Wilm, A., Lopez, R.*, et al.* (2007). Clustal W and Clustal X version 2.0. Bioinformatics *23*, 2947-2948.

Ley, R.E., Hamady, M., Lozupone, C., Turnbaugh, P.J., Ramey, R.R., Bircher, J.S., Schlegel, M.L., Tucker, T.A., Schrenzel, M.D., Knight, R.*, et al.* (2008). Evolution of mammals and their gut microbes. Science (New York, NY) *320*, 1647-1651.

Li, G., Badaluddin, N.A., and Kitakawa, M. (2015). Characterization of inner membrane protein YciB in Escherichia coli: YciB interacts with cell elongation and division proteins. Microbiol Immunol *59*, 700-704.

Liberati, N.T., Urbach, J.M., Miyata, S., Lee, D.G., Drenkard, E., Wu, G., Villanueva, J., Wei, T., and Ausubel, F.M. (2006). An ordered, nonredundant library of *Pseudomonas aeruginosa* strain PA14 transposon insertion mutants. Proc Natl Acad Sci U S A *103*, 2833-2838.

Liu, J.D., and Parkinson, J.S. (1989). Genetics and sequence analysis of the pcnB locus, an Escherichia coli gene involved in plasmid copy number control. J Bacteriol *171*, 1254-1261.

Loeb, J.D., Sepulveda-Becerra, M., Hazan, I., and Liu, H. (1999). A G1 cyclin is necessary for maintenance of filamentous growth in Candida albicans. Mol Cell Biol *19*, 4019-4027.

Luo, Q., Pagel, P., Vilne, B., and Frishman, D. (2011). DIMA 3.0: Domain interaction map. Nucleic Acids Research *39*.

Lupoli, T.J., Lebar, M.D., Markovski, M., Bernhardt, T., Kahne, D., and Walker, S. (2014). Lipoprotein activators stimulate Escherichia coli penicillin-binding proteins by different mechanisms. Journal of the American Chemical Society *136*, 52-55.

Madsen, J.S., Lin, Y.-C., Squyres, G.R., Price-Whelan, A., de Santiago Torio, A., Song, A., Cornell, W.C., Sørensen, S.J., Xavier, J.B., and Dietrich, L.E.P. (2015). Facultative control of matrix production optimizes competitive fitness in Pseudomonas aeruginosa PA14 biofilm models. Applied and Environmental Microbiology *81*, AEM.02628-02615.

Manning, C.D., Sch\, H., \#252, and tze (1999). Foundations of statistical natural language processing. 680.

Marbach, A., and Bettenbrock, K. (2012). lac operon induction in Escherichia coli: Systematic comparison of IPTG and TMG induction and influence of the transacetylase LacA. J Biotechnol *157*, 82-88.

Markovski, M., Bohrhunter, J.L., Lupoli, T.J., Uehara, T., Walker, S., Kahne, D.E., and Bernhardt, T.G. (2016). Cofactor bypass variants reveal a conformational control mechanism governing cell wall polymerase activity. Proceedings of the National Academy of Sciences *113*, 4788-4793.

Marr A N, Y.D., and Hildreth, D.E. (1980). Theory of edge detection. Proc R Boc Lond B *207*, 187-217.

Masi, M., and Wandersman, C. (2010). Multiple signals direct the assembly and function of a type 1 secretion system. Journal of Bacteriology *192*, 3861-3869.

Matsuhashi, M., Maruyama, I.N., Takagaki, Y., Tamaki, S., Nishimura, Y., and Hirota, Y. (1978). Isolation of a mutant of Escherichia coli lacking penicillin-sensitive D-alanine carboxypeptidase IA. Proceedings of the National Academy of Sciences of the United States of America *75*, 2631-2635.

Meeske, A.J., Rodrigues, C.D., Brady, J., Lim, H.C., Bernhardt, T.G., and Rudner, D.Z. (2016). High-Throughput Genetic Screens Identify a Large and Diverse Collection of New Sporulation Genes in Bacillus subtilis. PLoS Biol *14*, e1002341.

Memarian, N., Jessulat, M., Alirezaie, J., Mir-Rashed, N., Xu, J., Zareie, M., Smith, M., Golshani, A., Hughes, T.R., Robinson, M.D., *et al.* (2007). Colony size measurement of the yeast gene deletion strains for functional genomics. BMC Bioinformatics *8*, 117-117.

Memišević, V., Wallqvist, A., Reifman, J., Hart, G., Ramani, A., Marcotte, E., Sambourg, L., Thierry-Mieg, N., Stumpf, M., Thorne, T., *et al.* (2013). Reconstituting protein interaction networks using parameter-dependent domain-domain interactions. BMC Bioinformatics *14*, 154.

Mende, D.R., Sunagawa, S., Zeller, G., and Bork, P. (2013). Accurate and universal delineation of prokaryotic species. Nat Methods *10*, 881-884.

Merritt, J.H., Brothers, K.M., Kuchma, S.L., and O'Toole, G.A. (2007). SadC reciprocally influences biofilm formation and swarming motility via modulation of exopolysaccharide production and flagellar function. J Bacteriol *189*, 8154-8164.

Miller, J.H. (1972). Experiments in molecular genetics. Cold spring harbor, NY *433*, 352-355.

Miquel, S., Martín, R., Rossi, O., Bermúdez-Humarán, L.G., Chatel, J.M., Sokol, H., Thomas, M., Wells, J.M., and Langella, P. (2013). Faecalibacterium prausnitzii and human intestinal health. In Current Opinion in Microbiology, pp. 255-261.

Mitchell, A.P. (1998). Dimorphism and virulence in Candida albicans. Current Opinion in Microbiology *1*, 687-692.

Moscoso, J.A., Jaeger, T., Valentini, M., Hui, K., Jenal, U., and Filloux, A. (2014). The diguanylate cyclase SadC is a central player in Gac/Rsm-mediated biofilm formation in Pseudomonas aeruginosa. J Bacteriol *196*, 4081-4088.

Murad, A.M., Leng, P., Straffon, M., Wishart, J., Macaskill, S., MacCallum, D., Schnell, N., Talibi, D., Marechal, D., Tekaia, F., *et al.* (2001). NRG1 represses yeast-hypha morphogenesis and hypha-specific gene expression in Candida albicans. EMBO J *20*, 4742-4752.

Musken, M., Di Fiore, S., Dotsch, A., Fischer, R., and Haussler, S. (2010). Genetic determinants of Pseudomonas aeruginosa biofilm establishment. Microbiology *156*, 431-441.

Nadal Jimenez, P., Koch, G., Thompson, J.A., Xavier, K.B., Cool, R.H., Quax, W.J., Jimenez, P.N., Koch, G., Thompson, J.A., Xavier, K.B., *et al.* (2012). The multiple signaling systems regulating virulence in Pseudomonas aeruginosa. Microbiol Mol Biol Rev *76*, 46-65.

Navarro, M.V.A.S., De, N., Bae, N., Wang, Q., and Sondermann, H. (2009). Structural analysis of the GGDEF-EAL domain-containing c-di-GMP receptor FimX. Structure (London, England : 1993) *17*, 1104-1116.

Nichols, R.J., Sen, S., Choo, Y.J., Beltrao, P., Zietek, M., Chaba, R., Lee, S., Kazmierczak, K.M., Lee, K.J., Wong, A., *et al.* (2011). Phenotypic landscape of a bacterial cell. Cell *144*, 143-156.

Nikaido, H. (2003). Molecular basis of bacterial outer membrane permeability revisited. Microbiol Mol Biol Rev *67*, 593-656.

Nobile, C.J., Fox, E.P., Nett, J.E., Sorrells, T.R., Mitrovich, Q.M., Hernday, A.D., Tuch, B.B., Andes, D.R., and Johnson, A.D. (2012). A recently evolved transcriptional network controls biofilm development in Candida albicans. Cell *148*, 126-138.

Noble, S.M., French, S., Kohn, L.A., Chen, V., and Johnson, A.D. (2010). Systematic screens of a Candida albicans homozygous deletion library decouple morphogenetic switching and pathogenicity. Nat Genet *42*, 590-598.

Noble, S.M., and Johnson, A.D. (2005). Strains and strategies for large-scale gene deletion studies of the diploid human fungal pathogen Candida albicans. Eukaryot Cell *4*, 298-309.

Ochoa, D., and Pazos, F. (2014). Practical aspects of protein co-evolution. Frontiers in cell and developmental biology *2*, 14.

Ohya, Y., Sese, J., Yukawa, M., Sano, F., Nakatani, Y., Saito, T.L., Saka, A., Fukuda, T., Ishihara, S., Oka, S., *et al.* (2005). High-dimensional and large-scale phenotyping of yeast mutants. Proceedings of the National Academy of Sciences *102*, 19015-19020.

Otsu, N. (1979). A threshold selection method from gray-level histograms. IEEE Transactions on Systems, Man, and Cybernetics *9*, 62-66.

Pagel, P., Oesterheld, M., Stümpflen, V., and Frishman, D. (2006). The DIMA web resourceexploring the protein domain network. Bioinformatics *22*, 997-998.

Pagel, P., Oesterheld, M., Tovstukhina, O., Strack, N., Stümpflen, V., and Frishman, D. (2008). DIMA 2.0--predicted and known domain interactions. Nucleic acids research *36*, D651-655.

Paradis-Bleau, C., Kritikos, G., Orlova, K., Typas, A., and Bernhardt, T.G. (2014). A genome-wide screen for bacterial envelope biogenesis mutants identifies a novel factor involved in cell wall precursor metabolism. PLoS Genet *10*, e1004056.

Paradis-Bleau, C., Markovski, M., Uehara, T., Lupoli, T.J., Walker, S., Kahne, D.E., and Bernhardt, T.G. (2010). Lipoprotein cofactors located in the outer membrane activate bacterial cell wall polymerases. Cell *143*, 1110-1120.

Park, J.T., and Uehara, T. (2008). How bacteria consume their own exoskeletons (turnover and recycling of cell wall peptidoglycan). Microbiol Mol Biol Rev *72*, 211-227.

Parsons, A.B., Lopez, A., Givoni, I.E., Williams, D.E., Gray, C.A., Porter, J., Chua, G., Sopko, R., Brost, R.L., Ho, C.H., *et al.* (2006). Exploring the mode-of-action of bioactive compounds by chemical-genetic profiling in yeast. Cell *126*, 611-625.

Pasquina, L., Santa Maria, J.P., McKay Wood, B., Moussa, S.H., Matano, L.M., Santiago, M., Martin, S.E.S., Lee, W., Meredith, T.C., and Walker, S. (2016). A synthetic lethal approach for compound and target identification in Staphylococcus aureus. Nature chemical biology *12*, 40-45.

Pazos, F., Juan, D., Izarzugaza, J.M., Leon, E., and Valencia, A. (2008). Prediction of protein interaction based on similarity of phylogenetic trees. Methods Mol Biol *484*, 523-535.

Pazos, F., Ranea, J.A.G., Juan, D., and Sternberg, M.J.E. (2005). Assessing protein co-evolution in the context of the tree of life assists in the prediction of the interactome. J Mol Biol *352*, 1002-1015.

Pazos, F., and Valencia, A. (2001). Similarity of phylogenetic trees as indicator of protein-protein interaction. Protein engineering *14*, 609-614.

Pennartz, A., Généreux, C., Parquet, C., Mengin-Lecreulx, D., and Joris, B. (2009). Substrate-induced inactivation of the Escherichia coli AmiD N-acetylmuramoyl-L-alanine amidase highlights a new strategy to inhibit this class of enzyme. Antimicrobial agents and chemotherapy *53*, 2991-2997.

Phillips, John W.W., Goetz, Michael A.A., Smith, Scott K.K., Zink, Deborah L.L., Polishook, J., Onishi, R., Salowe, S., Wiltsie, J., Allocco, J., Sigmund, J., *et al.* (2011). Discovery of kibdelomycin, a potent new class of bacterial type II topoisomerase inhibitor by chemical-genetic profiling in Staphylococcus aureus. Chem Biol *18*, 955-965.

Pierson, L.S., and Pierson, E.A. (2010). Metabolism and function of phenazines in bacteria: impacts on the behavior of bacteria in the environment and biotechnological processes. Applied microbiology and biotechnology *86*, 1659-1670.

Popescu, A.-M., and Etzioni, O. (2007). Extracting product features and opinions from reviews. In Natural language processing and text mining (Springer), pp. 9-28.

Porwollik, S., Santiviago, C.A., Cheng, P., Long, F., Desai, P., Fredlund, J., Srikumar, S., Silva, C.A., Chu, W., Chen, X., *et al.* (2014). Defined single-gene and multi-gene deletion mutant collections in salmonella enterica sv typhimurium. PLoS ONE *9*, e99820.

Powell, S., Forslund, K., Szklarczyk, D., Trachana, K., Roth, A., Huerta-Cepas, J., Gabaldón, T., Rattei, T., Creevey, C., Kuhn, M., *et al.* (2014). EggNOG v4.0: Nested orthology inference across 3686 organisms. Nucleic Acids Research *42*.

Prats, R., and de Pedro, M.A. (1989). Normal growth and division of Escherichia coli with a reduced amount of murein. J Bacteriol *171*, 3740-3745.

Prigent-Combaret, C., Prensier, G., Le Thi, T.T., Vidal, O., Lejeune, P., and Dorel, C. (2000). Developmental pathway for biofilm formation in curli-producing Escherichia coli strains: role of flagella, curli and colanic acid. Environ Microbiol *2*, 450-464.

Qi, Y., Chuah, M.L.C., Dong, X., Xie, K., Luo, Z., Tang, K., and Liang, Z.-X. (2011). Binding of cyclic diguanylate in the non-catalytic EAL domain of FimX induces a long-range conformational change. The Journal of biological chemistry *286*, 2910-2917.

Reynolds, K.A., McLaughlin, R.N., and Ranganathan, R. (2011). Hot spots for allosteric regulation on protein surfaces. Cell *147*, 1564-1575.

Rice, K.C., and Bayles, K.W. (2008). Molecular control of bacterial death and lysis. Microbiology and molecular biology reviews : MMBR *72*, 85-109, table of contents.

Roemer, T., Davies, J., Giaever, G., and Nislow, C. (2012). Bugs, drugs and chemical genomics. Nat Chem Biol *8*, 46-56.

Romantsov, T., Guan, Z., and Wood, J.M. (2009). Cardiolipin and the osmotic stress responses of bacteria. Biochimica et biophysica acta *1788*, 2092-2100.

Rost, B. (2001). Review: protein secondary structure prediction continues to rise. Journal of structural biology *134*, 204-218.

Ryan, C.J., Cimermančič, P., Szpiech, Z.A., Sali, A., Hernandez, R.D., and Krogan, N.J. (2013). High-resolution network biology: connecting sequence with function. Nature Reviews Genetics *14*, 865-879.

Ryan, O., Shapiro, R.S., Kurat, C.F., Mayhew, D., Baryshnikova, A., Chin, B., Lin, Z.-Y.Y., Cox, M.J., Vizeacoumar, F.J., Cheung, D., *et al.* (2012). Global gene deletion analysis exploring yeast filamentous growth. Science *337*, 1353-1356.

Saitou, N., and Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. Molecular Biology and Evolution *4*, 406-425.

Saka, K., Tadenuma, M., Nakade, S., Tanaka, N., Sugawara, H., Nishikawa, K., Ichiyoshi, N., Kitagawa, M., Mori, H., Ogasawara, N., *et al.* (2005). A complete set of Escherichia coli open reading frames in mobile plasmids facilitating genetic studies. DNA Res *12*, 63-68.

Sato, T., Yamanishi, Y., Kanehisa, M., and Toh, H. (2005). The inference of protein-protein interactions by co-evolutionary analysis is improved by excluding the information about the phylogenetic relationships. Bioinformatics *21*, 3482-3489.

Sauvage, E., Kerff, F., Terrak, M., Ayala, J.A., and Charlier, P. (2008). The penicillin-binding proteins: structure and role in peptidoglycan biosynthesis. FEMS Microbiol Rev *32*, 234-258.

Scheurwater, E.M., and Burrows, L.L. (2011). Maintaining network security: How macromolecular structures cross the peptidoglycan layer. FEMS Microbiology Letters *318*, 1-9.

Schiffer, G., and Holtje, J.V. (1999). Cloning and characterization of PBP 1C, a third member of the multimodular class A penicillin-binding proteins of Escherichia coli. J Biol Chem *274*, 32031-32039.

Schleifer, K.H., and Kandler, O. (1972). Peptidoglycan types of bacterial cell walls and their taxonomic implications. Bacteriological reviews *36*, 407-477.

Schuldiner, M., Collins, S.R., Thompson, N.J., Denic, V., Bhamidipati, A., Punna, T., Ihmels, J., Andrews, B., Boone, C., Greenblatt, J.F., *et al.* (2005). Exploration of the function and organization of the yeast early secretory pathway through an epistatic miniarray profile. Cell *123*, 507-519.

Schuldiner, M., Collins, S.R., Weissman, J.S., and Krogan, N.J. (2006). Quantitative genetic analysis in Saccharomyces cerevisiae using epistatic miniarray profiles (E-MAPs) and its application to chromatin functions. Methods *40*, 344-352.

Schweizer, A., Rupp, S., Taylor, B.N., Rollinghoff, M., and Schroppel, K. (2000). The TEA/ATTS transcription factor CaTec1p regulates hyphal development and virulence in Candida albicans. Mol Microbiol *38*, 435-445.

Sellam, A., Askew, C., Epp, E., Tebbji, F., Mullick, A., Whiteway, M., and Nantel, A. (2010). Role of transcription factor CaNdt80p in cell separation, hyphal growth, and virulence in Candida albicans. Eukaryot Cell *9*, 634-644.

Serra, D.O., Richter, A.M., and Hengge, R. (2013). Cellulose as an architectural element in spatially structured Escherichia coli biofilms. Journal of bacteriology *195*, 5540-5554.

Setlow, P. (2003). Spore germination. Current opinion in microbiology *6*, 550-556.

Shiver, A.L., Osadnik, H., Kritikos, G., Li, B., Krogan, N., Typas, A., and Gross, C.A. (2016). A Chemical-Genomic Screen of Neglected Antibiotics Reveals Illicit Transport of Kasugamycin and Blasticidin S. PLoS Genet *12*, e1006124.

Silhavy, T.J., Kahne, D., and Walker, S. (2010). The bacterial cell envelope. Cold Spring Harbor perspectives in biology *2*, a000414-a000414.

Silhavy, T.J., Ruiz, N., Kahne, D., Silhavy, T.J., Ruiz, N., and Kahne, D. (2006). Advances in understanding bacterial outer-membrane biogenesis. Nat Rev Microbiol *4*, 57-66.

Simm, R., Ahmad, I., Rhen, M., Le Guyon, S., and Romling, U. (2014). Regulation of biofilm formation in Salmonella enterica serovar Typhimurium. Future Microbiol *9*, 1261-1282.

Simm, R., Morr, M., Kader, A., Nimtz, M., and Römling, U. (2004). GGDEF and EAL domains inversely regulate cyclic di-GMP levels and transition from sessility to motility. Molecular microbiology *53*, 1123-1134.

Singh, S., Carpenter, A.E., and Genovesio, A. (2014). Increasing the Content of High-Content Screening: An Overview. Journal of Biomolecular Screening *19*, 640-650.

Singh, S.K., SaiSree, L., Amrutha, R.N., and Reddy, M. (2012). Three redundant murein endopeptidases catalyse an essential cleavage step in peptidoglycan synthesis of Escherichia coli K12. Molecular Microbiology *86*, 1036-1051.

Small, I., Peeters, N., Legeai, F., and Lurin, C. (2004). Predotar: A tool for rapidly screening proteomes for N-terminal targeting sequences. Proteomics *4*, 1581-1590.

Smith, A.R. (1978). Color gamut transform pairs. ACM SIGGRAPH Computer Graphics *12*, 12-19.

Starkey, M., Hickman, J.H., Ma, L., Zhang, N., De Long, S., Hinz, A., Palacios, S., Manoil, C., Kirisits, M.J., Starner, T.D., *et al.* (2009). Pseudomonas aeruginosa rugose small-colony variants have adaptations that likely promote persistence in the cystic fibrosis lung. Journal of bacteriology *191*, 3492-3503.

Stoldt, V.R., Sonneborn, A., Leuker, C.E., and Ernst, J.F. (1997). Efg1p, an essential regulator of morphogenesis of the human pathogen Candida albicans, is a member of a conserved class of bHLH proteins regulating morphogenetic processes in fungi. EMBO J *16*, 1982-1991.

Strimmer, K. (2008). fdrtool: A versatile R package for estimating local and tail area-based false discovery rates. Bioinformatics *24*, 1461-1462.

Sturges, W.S., and Rettger, L.F. (1922). Bacterial Autolysis. J Bacteriol *7*, 551-577.

Styles, E.B., Friesen, H., Boone, C., and Andrews, B.J. (2016). High-Throughput Microscopy-Based Screening in Saccharomyces cerevisiae. Cold Spring Harbor Protocols *2016*, pdb.top087593.

Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., *et al.* (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A *102*, 15545-15550.

Sudbery, P., Gow, N., and Berman, J. (2004). The distinct morphogenic states of Candida albicans. Trends Microbiol *12*, 317-324.

Sun, J., Li, Y., and Zhao, Z. (2007). Phylogenetic profiles for the prediction of protein-protein interactions: How to select reference organisms? Biochemical and Biophysical Research Communications *353*, 985-991.

Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., Simonovic, M., Roth, A., Santos, A., and Tsafou, K.P. (2014). STRING v10: protein–protein interaction networks, integrated over the tree of life. Nucl Acids Res, gku1003.

Takeuchi, R., Tamura, T., Nakayashiki, T., Tanaka, Y., Muto, A., Wanner, B.L., Mori, H., Babu, M., Diaz-Mejia, J., Vlasblom, J., *et al.* (2014). Colony-live —a high-throughput method for measuring microbial colony growth kinetics— reveals diverse growth effects of gene knockouts in Escherichia coli. BMC Microbiology *14*, 171.

Tatar, L.D., Marolda, C.L., Polischuk, A.N., van Leeuwen, D., and Valvano, M.A. (2007). An Escherichia coli undecaprenyl-pyrophosphate phosphatase implicated in undecaprenyl phosphate recycling. Microbiology *153*, 2518-2529.

Tatusov, R.L., Galperin, M.Y., Natale, D.A., and Koonin, E.V. (2000). The COG database: a tool for genome-scale analysis of protein functions and evolution. Nucleic acids research *28*, 33-36.

Tatusov, R.L., Koonin, E.V., and Lipman, D.J. (1997). A genomic perspective on protein families. Science (New York, NY) *278*, 631-637.

Team, R.C. (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2013 (ISBN 3-900051-07-0).

Team, R.D.C. (2015). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing *1*, 409.

Thompson, J.D., Higgins, D.G., and Gibson, T.J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic acids research *22*, 4673-4680.

Tillier, E.R.M., and Charlebois, R.L. (2009). The human protein coevolution network. Genome Research *19*, 1861-1871.

Tong, A.H.Y., Lesage, G., Bader, G.D., Ding, H., Xu, H., Xin, X., Young, J., Berriz, G.F., Brost, R.L., Chang, M., *et al.* (2004). Global mapping of the yeast genetic interaction network. Science *303*, 808-813.

Turnbaugh, P.J., Hamady, M., Yatsunenko, T., Cantarel, B.L., Ley, R.E., Sogin, M.L., Jones, W.J., Roe, B.a., Jason, P., Egholm, M., *et al.* (2009). A core gut microbiom in obese and lean twins. Nature *457*, 480-484.

Turner, R.D., Vollmer, W., and Foster, S.J. (2014). Different walls for rods and balls: the diversity of peptidoglycan. Molecular Microbiology *91*, 862-874.

Typas, A., Banzhaf, M., Gross, C.A., and Vollmer, W. (2012). From the regulation of peptidoglycan synthesis to bacterial growth and morphology. Nature reviews Microbiology *10*, 123-136.

Typas, A., Banzhaf, M., van den Berg van Saparoea, B., Verheul, J., Biboy, J., Nichols, R.J., Zietek, M., Beilharz, K., Kannenberg, K., von Rechenberg, M., *et al.* (2010). Regulation of peptidoglycan synthesis by outer-membrane proteins. Cell *143*, 1097-1109.

Typas, A., Nichols, R.J., Siegele, D.A., Shales, M., Lim, B., Braberg, H., Yamamoto, N., Takeuchi, R., Barry, L., Mori, H., *et al.* (2008). High-throughput, quantitative analyses of genetic interactions in E. coli. Nat Methods *5*, 781-787.

Uehara, T., Parzych, K.R., Dinh, T., and Bernhardt, T.G. (2010). Daughter cell separation is controlled by cytokinetic ring-activated cell wall hydrolysis. EMBO J *29*, 1412-1422.

Van Houdt, R., and Michiels, C.W. (2005). Role of bacterial cell surface structures in Escherichia coli biofilm formation. Res Microbiol *156*, 626-633.

Van Rijsbergen, C.J. (1979). Information Retrieval, 2nd edition. Butterworths.

Vertommen, D., Ruiz, N., Leverrier, P., Silhavy, T.J., and Collet, J.-F. (2009). Characterization of the role of the Escherichia coli periplasmic chaperone SurA using differential proteomics. Proteomics *9*, 2432-2443.

Vollmer, W., and Seligman, S.J. (2010). Architecture of peptidoglycan: more data and more models. Trends Microbiol *18*, 59-66.

Vollmer, W., von Rechenberg, M., and Höltje, J.-V. (1999). Demonstration of molecular interactions between the murein polymerase PBP1B, the lytic transglycosylase MltA, and the scaffolding protein MipA of Escherichia coli. J Biol Chem *274*, 6726-6734.

von Mering, C., Huynen, M., Jaeggi, D., Schmidt, S., Bork, P., and Snel, B. (2003). STRING: a database of predicted functional associations between proteins. Nucleic acids research *31*, 258-261.

Wagih, O., and Parts, L. (2014). Gitter: a Robust and Accurate Method for Quantification of Colony Sizes From Plate Images. G3 (Bethesda, Md) *4*, 547-552.

Wang, X., and Quinn, P.J. (2010). Lipopolysaccharide: Biosynthetic pathway and structure modification. Progress in Lipid Research *49*, 97-107.

Wang, Y., Wilks, J.C., Danhorn, T., Ramos, I., Croal, L., and Newman, D.K. (2011). Phenazine-1-Carboxylic Acid Promotes Bacterial Biofilm Development via Ferrous Iron Acquisition. Journal of Bacteriology *193*, 3606-3617.

Wetherow, O., Green, J., Chan, A.D.C., and Golshani, A. (2010). Plate analyzer - a yeast colony size measurement system (IEEE).

Xu, P., Ge, X., Chen, L., Wang, X., Dou, Y., Xu, J.Z., Patel, J.R., Stone, V., Trinh, M., Evans, K., *et al.* (2011). Genome-wide essential gene identification in Streptococcus sanguinis. Scientific reports *1*, 125.

Yahashiri, A., Jorgenson, M.A., and Weiss, D.S. (2015). Bacterial SPOR domains are recruited to septal peptidoglycan by binding to glycan strands that lack stem peptides. Proceedings of the National Academy of Sciences of the United States of America *112*, 11347-11352.

Yang, X.-B., Jin, X.-Q., Du, Z.-M., and Zhu, Y.-H. (2011). A novel model-based fault detection method for temperature sensor using fractal correlation dimension. Building and Environment *46*, 970-979.

Zahrl, D., Wagner, M., Bischof, K., Bayer, M., Zavecz, B., Beranek, A., Ruckenstuhl, C., Zarfel, G.E., and Koraimann, G. (2005). Peptidoglycan degradation by specialized lytic transglycosylases associated with type III and type IV secretion systems. Microbiology *151*, 3455-3467.

Zanella, F., Lorens, J.B., and Link, W. (2010). High content screening: seeing is believing. Trends in Biotechnology *28*, 237-245.

Zeghouf, M., Li, J., Butland, G., Borkowska, A., Canadien, V., Richards, D., Beattie, B., Emili, A., and Greenblatt, J.F. (2004). Sequential Peptide Affinity (SPA) system for the identification of mammalian and bacterial protein complexes. J Proteome Res *3*, 463-468.

Zeytuni, N., Zarivach, R., Bahadur, R.P., Zacharias, M., Baker, M.J., Frazier, A.E., Gulbis, J.M., Ryan, M.T., Biegert, A., Mayer, C., *et al.* (2012). Structural and functional discussion of the tetra-trico-peptide repeat, a protein interaction module. Structure *20*, 397-405.

Zogaj, X., Bokranz, W., Nimtz, M., and Römling, U. (2003). Production of cellulose and curli fimbriae by members of the family Enterobacteriaceae isolated from the human gastrointestinal tract. Infection and immunity *71*, 4151-4158.

Zogaj, X., Nimtz, M., Rohde, M., Bokranz, W., Romling, U., and Römling, U. (2001). The multicellular morphotypes of Salmonella typhimurium and Escherichia coli produce cellulose as the second component of the extracellular matrix. Mol Microbiol *39*, 1452-1463.

Publications

A. A Genome-Wide Screen for Bacterial Envelope Biogenesis Mutants Identifies a Novel Factor Involved in Cell Wall Precursor Metabolism

Published in PLoS Genetics 02.01.2014

Citation:
Paradis-Bleau, C., Kritikos, G., Orlova, K., Typas, A., and Bernhardt, T.G. (2014). A genome-wide screen for bacterial envelope biogenesis mutants identifies a novel factor involved in cell wall precursor metabolism. PLoS Genet 10, e1004056.

The manuscript is not included in the printed version of this thesis. Interested readers are welcome to the published manuscript, available at the following link: http://goo.gl/gvyt2N

B. A Chemical-Genomic Screen of Neglected Antibiotics Reveals Illicit Transport of Kasugamycin and Blasticidin S

Published in PLoS Genetics 29.06.2016

Citation:
Shiver, A.L., Osadnik, H., Kritikos, G., Li, B., Krogan, N., Typas, A., and Gross, C.A. (2016). A Chemical-Genomic Screen of Neglected Antibiotics Reveals Illicit Transport of Kasugamycin and Blasticidin S. PLoS Genet 12, e1006124.

The manuscript is not included in the printed version of this thesis. Interested readers are welcome to the published manuscript, available at the following link: http://goo.gl/hRE9ow

Manuscripts submitted for publication or in preparation

A. Building Systems Resources for the Model Gram-positive bacterium *Bacillus subtilis*

Submitted for publication in Cell Systems, currently under review.

The manuscript is not included in the printed version of this thesis. Interested readers are welcome to the submitted version of the manuscript, available at the following link:
https://goo.gl/UJoM26

B. Iris: expanding the palette of microbial phenotypic readouts

Manuscript submitted to Nature Microbiology, currently under review.

The manuscript is not included in the printed version of this thesis. Interested readers are welcome to the submitted version of the manuscript, available at the following link:
https://goo.gl/qzk2sK