

Mathematical Modeling and Analysis of Gene Expression to Understand Phenotypic
Heterogeneity and the Response of *Methylobacterium extorquens* to Formaldehyde Toxicity

A Dissertation

Presented in Partial Fulfillment of the Requirements for the
Degree of Doctorate of Philosophy

with a

Major in in Bioinformatics and Computational Biology

in the

College of Graduate Studies

University of Idaho

by

Siavash Riazzi

Major Professors: Christopher Marx, Ph.D.; Christopher Remien, Ph.D.

Committee Members: Holly Wichman, Ph.D.; Steve Krone, Ph.D.; Craig Miller, Ph.D.

Department Administrator: David Tank, Ph.D.

December 2018

Authorization to Submit Dissertation

This dissertation of Siavash Riazi, submitted for the degree of Doctorate of Philosophy with a Major in Bioinformatics and Computational Biology and titled "Mathematical Modeling and Analysis of Gene Expression to Understand Phenotypic Heterogeneity and the Response of *Methylobacterium extorquens* to Formaldehyde Toxicity," has been reviewed in final form. Permission, as indicated by the signatures and dates below, is now granted to submit final copies to the College of Graduate Studies for approval.

Major Professors: _____ Date: _____
Christopher Marx, Ph.D.

_____ Date: _____
Christopher Remien, Ph.D.

Committee Members: _____ Date: _____
Holly Wichman, Ph.D.

_____ Date: _____
Steve Krone, Ph.D.

_____ Date: _____
Craig Miller, Ph.D.

Department
Administrator: _____ Date: _____
David Tank, Ph.D.

Abstract

Methylobacterium extorquens is a facultative methylotrophic bacterium that lives on plant leaves. As part of the natural oxidation pathway of methanol secreted from the leaves, formaldehyde is generated. Experiments have shown there is phenotypic heterogeneity in tolerance to formaldehyde, and this heterogeneity varies continuously. Exposing *M. extorquens* to a high concentration (4 mM) of formaldehyde changed the distribution of tolerance to formaldehyde. In the second chapter of this dissertation, I introduced a mathematical model to investigate the processes involved in the change of the tolerance distribution. The model suggests there is an absolute threshold between survival and death in face of the stress from formaldehyde. In addition, I showed growth and death are not sufficient to explain the change of distribution of tolerance, and in fact, there is also a suggestion for phenotypic movements that permit the cells to change their phenotypic states. Moreover, the model showed that the phenotypic movements that occur depend upon the environmental conditions. In the third chapter, I investigated the genes involved in response to formaldehyde stress using RNA-seq analysis. In order to find specific mechanisms involved in formaldehyde-mediated translation inhibition, cultures of bacteria treated with either formaldehyde and kanamycin was investigated. To assess the role of the EfgA protein – which has a role in translation inhibition by formaldehyde – WT and $\Delta efgA$ strains were investigated in the mentioned treatments. I showed that a great portion of the response to formaldehyde is shared with the kanamycin response, and that having EfgA protein is crucial to the formaldehyde stress response. Analysis of functional gene groups showed that cytochromes, chaperones, DNA damage repair system, ABC transporters and flagellar proteins are among the highly affected genes in response to the formaldehyde. This analysis of RNA-seq data provides a set of candidate genes that potentially have role in the phenotypic heterogeneity in tolerance to formaldehyde.

Acknowledgements

First I would like to appreciate my supervisors for all of their support during graduate school: Christopher Marx, It's hard to sum up a five year journey in a few words. We came a long way together, from starting an experimental project to a mixture of experiment and modeling and finally what I'm presenting in this dissertation. Chris taught me how to think like a biologist, question is always first, then methods and modeling! Thanks Chris for all the great times I had with you, in your house, in Moscow pubs, over trips at conferences and elsewhere during the past few years!

Christopher Remien, it was fall semester of 2015 when I decided to do my lab rotation with Chris Remien, I started to work on modeling formaldehyde toxicity. This project kept me involved for the next 3 years and Chris became my second advisor. He taught me how to think like a mathematician: simplicity is always the key, adding the details doesn't make things necessarily better! Thanks Chris, I will always try to keep this mindset in my future work!

I want to acknowledge Ben Ridenhour for all his help on technical parts of the project. We had many long meetings in Ben's office that were supposed to be for an hour, but lasted three hours, from working on probability distributions to debugging tedious STAN codes. Without Ben, I would not have been able to elaborate this project with such a level of complexity. Thanks Ben!

I want to thank my experimental collaborators:

Jessica Lee for providing experimental data of heterogeneity project. Jessica was not only a fantastic colleague but she also taught me basic R programming, plus all the great cakes that made our parties sweeter; thanks Jessica for everything!

And Jannell Bazurto for providing experimental data of RNA-seq project. I know all her help to editing and revising this document. Jannell was one of my best friends in the US and will be missed deeply!

And I would like to appreciate my committee members for all their advice and feedback on this project:

Holly Wichman: she always reminded me BCB deadlines and regulations, thanks for keeping me on track!

Steve Krone: First, Steve is a great teacher, thanks for Mathematical Biology, Mathematical Genetics and Stochastic Methods, great courses! Steve's office was also always open, thanks for those consultations!

Craig Miller, Craig was supposed to serve on my committee for an experimental project, after changing projects he still accepted to serve on my committee, thanks Craig for all critical comments and feedbacks in meetings, I appreciate it!

And I have to acknowledge Eva Top, Eva was the director of BCB during most of my study in UI, always supportive and open to have a conversation, thanks Eva!

And finally, Lisha, throughout my graduate study I got the chance to meet many of the staff from different parts of the campus. I have to say: Lisha is an exception. Without her BCB could be a stressful experience, but I never felt any difficulties, because I knew someone was watching us all the time! Thanks Lisha to be in UI!

Dedication

تقدیم به خانواده خودم و خانواده همه دانشجویانی که از دیدن عزیزانشان محروم ماندند

Table of Contents

Authorization to Submit Dissertation.....	ii
Abstract	iii
Acknowledgements	iv
Dedication	vi
Table of Contents	vii
List of Figures	ix
List of Tables	xii
1 Introduction	1
2 Mathematical Modeling of Response to Formaldehyde in <i>Methylobacterium extorquens</i>	5
2.1 Introduction	5
2.2 Empirical model system and background data	7
2.3 Model development	14
2.4 Methods	18
2.4.1 Converting cumulative data to densities	18
2.4.2 Initial conditions	19
2.4.3 Parameter estimation	20
2.4.4 Death functions	22
2.4.5 Numerical solution of the PDE.....	23
2.4.6 Aggregating results of the continuous model to discrete bins	23
2.4.7 Likelihood ratio test and model selection	25
2.4.8 AIC calculation.....	25
2.5 Results	27
2.5.1 Bimodality of tolerance distribution during transition from net death to net growth suggests death is an absolute cutoff with tolerance level	27
2.5.2 Diffusion is sufficient to explain the tolerance shift in growth with formaldehyde ...	29
2.5.3 Advection is necessary to explain the shifting back of tolerance in regrowth on succinate medium, but not methanol medium	30
2.6 Discussion.....	32

2.7 Supplementary materials	34
2.7.1 Estimating parameters	34
3 Analyzing Gene Expression to Understand the Response of <i>M. extorquens</i> to the Toxicity of Formaldehyde.....	46
3.1 Introduction	46
3.2 Methods	49
3.2.1 Normalization of the data	49
3.2.2 Principal Component Analysis and heatmaps	49
3.2.3 Venn diagrams.....	50
3.2.4 Analysis of significance in candidate genes	50
3.3 Results	50
3.3.1 Overall pattern of gene expression changes revealed via Principal Component Analysis (PCA).....	50
3.3.2 Growth of WT and $\Delta efgA$ are similar to each other in the no-stressor condition.....	51
3.3.3 Treatment with kanamycin showed a delay in expression response compared to loss of viability.....	53
3.3.4 EfgA is the key component in response to formaldehyde	55
3.3.5 Response to kanamycin and formaldehyde involve shared pathways	57
3.3.6 Formaldehyde oxidation genes showed down-regulation in treatment with formaldehyde.....	58
3.3.7 Few loci with beneficial mutations during formaldehyde evolution showed a significant change in expression upon formaldehyde exposure	61
3.3.8 Formaldehyde-induced genes involved in response to other common stresses.....	61
3.3.9 The response of $\Delta efgA$ compared to WT involved general stress response proteins..	63
3.4 Discussion.....	64
3.5 Supplementary material.....	68
4 Conclusion.....	75
5 References.....	80
6 Appendix	87

List of Figures

Figure 1.1 - Relationship between different types of heterogeneity and the outcome of exposure to a stress.....	2
Figure 1.2 - Different gene expression profile between subpopulation is one of the possible mechanisms to observe heterogeneity.....	3
Figure 2.1 - Dynamics of persisters and sensitive cells (Balaban et al., 2004).....	6
Figure 2.2 - Dynamics of cells grown on different concentrations of formaldehyde.....	8
Figure 2.3 - Viability and measured formaldehyde data for growth on 4 mM formaldehyde	9
Figure 2.4 - Fluorescent membrane dye showed presence of a subpopulation in growth on formaldehyde	10
Figure 2.5 - Time-lapse microscopy showed bimodal (i.e., growth or non-growth) phenotypes in response to formaldehyde	11
Figure 2.6 - Continuous distribution of tolerance to formaldehyde	12
Figure 2.7 - Dynamic of tolerance distribution in growth on methanol with 4 mM formaldehyde....	13
Figure 2.8 - Change in tolerance distribution in regrowth on methanol or succinate	14
Figure 2.9 - Schematic diagram of growth and phenotypic changes over time	15
Figure 2.10 - Death function and how it affects the population's distribution.....	17
Figure 2.11 - Non-cumulative densities calculated from taking differences of cumulative densities.	19
Figure 2.12 - Extending the initial condition beyond the limit of detection	20
Figure 2.13 - Original distribution and extended distribution.....	20
Figure 2.14 - Death curves show an exponential decline in viability with increasing concentrations of formaldehyde	22
Figure 2.15 - Death rates fitted using the absolute death function.....	22
Figure 2.16 - Model result of change of tolerance distribution in growth with 4 mM formaldehyde, results from the absolute death version.....	24
Figure 2.17 - Model result of change of tolerance distribution in growth with 4 mM formaldehyde, results from relative death version	24
Figure 2.18 - Model results are aggregated into discrete bins to be compared with the data	28
Figure 2.19 - Comparison between the result of the model and the data at time 16 hours in linear scale	28
Figure 2.20 - Results of different death functions with diffusion included in the model.....	30
Figure 2.21 - Advection is necessary to capture the regrowth on succinate	32

Figure 2.22 - Fitting growth rate on 3.75 mM succinate with three replicates	35
Figure 2.23 - Residuals of succinate growth fit	35
Figure 2.24 - Fitting the growth rate on 15mM methanol.....	36
Figure 2.25 - Residuals of the methanol fit.....	36
Figure 2.26 - OD ₆₀₀ and number of viable cells show a linear relationship.....	37
Figure 2.27 - Growth on 3.75 mM succinate	38
Figure 2.28 - Simulated growth on 3.75mM succinate	38
Figure 2.29 - Simulated growth on 15mM methanol	40
Figure 2.30 - Growth on 15 mM methanol treated with 1 mM formaldehyde.....	41
Figure 2.31 - Formaldehyde measurement data and fit.....	42
Figure 2.32 - Viability and Nash assay data for growth on 4mM formaldehyde.....	42
Figure 2.33 - Formaldehyde measurement data (red) and simulated result using estimated V_{maxf} (dashed line).....	43
Figure 2.34 - Formaldehyde measurement (red) and simulated result using estimated V_{maxf} (dashed line)	43
Figure 3.1 - Common role of kanamycin and formaldehyde in translation inhibition and design of the stress exposure experiment	48
Figure 3.2 - Comparison of the growth response of WT (green) and $\Delta efgA$ mutant (red) in three different conditions: no-stressor, kanamycin and formaldehyde	48
Figure 3.3 - PCA plot of all treatments in all timepoints in both WT and $\Delta efgA$	51
Figure 3.4 - Temporal heatmap plot of average fold change of gene expression data in the growth on succinate with no-stressor, for WT and $\Delta efgA$ strains	52
Figure 3.5 - Venn diagrams of differentially expressed genes in WT and $\Delta efgA$ with no-stressor over time.....	53
Figure 3.6 - Temporal heatmap of both genotypes in treatment with kanamycin.....	54
Figure 3.7 - Venn diagrams of differentially expressed genes at 180 and 360 minutes post kanamycin treatment.....	54
Figure 3.8 - Venn diagram of up and down-regulated genes in the kanamycin treatment at 360 minutes	55
Figure 3.9 - Temporal heatmap plot of the two genotypes treated with formaldehyde	56
Figure 3.10 - Comparisons between 5 minutes and 20 minutes in treatment with formaldehyde	57
Figure 3.11 - Venn diagram of up/down regulated genes at 5 minutes with formaldehyde	57
Figure 3.12 - Comparison between up/down-regulated genes in formaldehyde at 5 minutes and kanamycin at 360 minutes in WT	58

Figure 3.13 - Expression changes in genes encoding enzymes involved in formaldehyde metabolism	60
Figure 4.1- A phenotypic distribution of tolerance could be correlated with different gene expression profiles in each subpopulation	77
Figure 4.2 - Different tolerance distributions to formaldehyde for <i>Cellulomonas fimi</i> , <i>Escherichia coli</i> , <i>Pseudomonas putida</i> , or <i>Methylobacterium extorquens</i>	78

List of Tables

Table 2.1 - Description of state variables and their units in the model.....	16
Table 2.2 - Different phenotypic movements and corresponding mathematical implementation.....	25
Table 2.3 - Parameters and their values	26
Table 2.4 - Different environmental conditions and corresponding population's response.....	34
Table 2.5 - Estimation of phenotypic movement parameters (diffusion and advection) for the growth in 4 mM formaldehyde scenario	44
Table 2.6 - Estimation of phenotypic movement parameters (diffusion and advection) for the re-growth scenario	45
Table 3.1 - Expression changes for genes involved in formaldehyde metabolism in both WT and $\Delta efgA$ at 5 minutes after exposure to formaldehyde.	59
Table 3.2 - Changes in expression upon formaldehyde exposure in genes that harbored beneficial mutations during formaldehyde evolution	61
Table 3.3 - Total number of annotated genes in functional gene groupings that were differentially expressed in WT treated with formaldehyde.....	63
Table 3.4 - P-values from t-tests of OD_{600} in WT and $\Delta efgA$ in kanamycin (left) and formaldehyde (right) compared to the no-stressor in different timepoint	68
Table 3.5 - Up and down-regulated genes in WT with no-stressor compared to WT pre-treatment from 5 minutes to 180 minutes.	69
Table 3.6 - Up and down-regulated genes in $\Delta efgA$ with no-stressor compared to $\Delta efgA$ pre-treatment from 5 minutes to 180 minutes.....	70
Table 3.7 - Genes that were up-regulated in $\Delta efgA$ at 5 minutes with formaldehyde compared to WT with formaldehyde at 5 minutes	70
Table 3.8 - Genes that were down-regulated in $\Delta efgA$ at 5 minutes with formaldehyde, compared to WT with formaldehyde at 5 minutes.....	74

1 Introduction

Not all genetically identical cells behave similarly. Phenotypic heterogeneity is defined as variation between isogenic individuals within a population. There are various mechanisms that can lead to phenotypic heterogeneity, such as the known phenomena of epigenetic modifications, like DNA methylation and histone modification, and post translation modifications that modulate protein activities; a number of other mechanisms have also been proposed such as cellular age, periodic oscillations and cell-to-cell interactions (Ackermann, 2015). The best characterized source of phenotypic heterogeneity is associated with noise in gene expression (Elowitz et al., 2002). In most cases, noise derived from bursts in transcription is negligible as the noise is averaged within multiple genes involved in a function (Kiviet et al., 2014). However, there are some cases where the noise in gene expression can make a demonstrable difference, especially in genes involved in stress response and metabolism, which are shown to have high variability in gene expression (Ackermann, 2015). The effect of noise in a system could be magnified in the presence of a feedback loop and result in a bistable system (Hasty et al., 2000; Thomas, 1981). Well-known examples of bimodal phenotypic heterogeneity in stress response and central carbon metabolism are bacterial persistence (Balaban et al., 2004), sporulation (Veening et al., 2008), and lactose utilization in *E. coli* (Choi et al., 2008).

Heterogeneity can help populations survive stressful conditions by providing variable outcomes in the face of a stressor. For example, if a phenotypically homogenous population (a population with small variations among individuals) encounters a lethal stressor in its environment, then presumably there is a uniform outcome: death of the entire population (Figure 1.1, panel A). But there are cases where the heterogeneity could be advantageous for a population. If variation among individuals is relatively broad and the stress is not strong enough to kill all the individuals, part of the population can survive the stressful condition. In a bimodal distribution of phenotypes (Figure 1.1, panel B), we observe heterogeneity as two discrete states, even though not all sensitive cells have the same gene regulation profile (e.g. persistence). In a wider distribution of phenotypes across a population, we can observe multiple subpopulations surviving the stress (Figure 1.1, panel C).

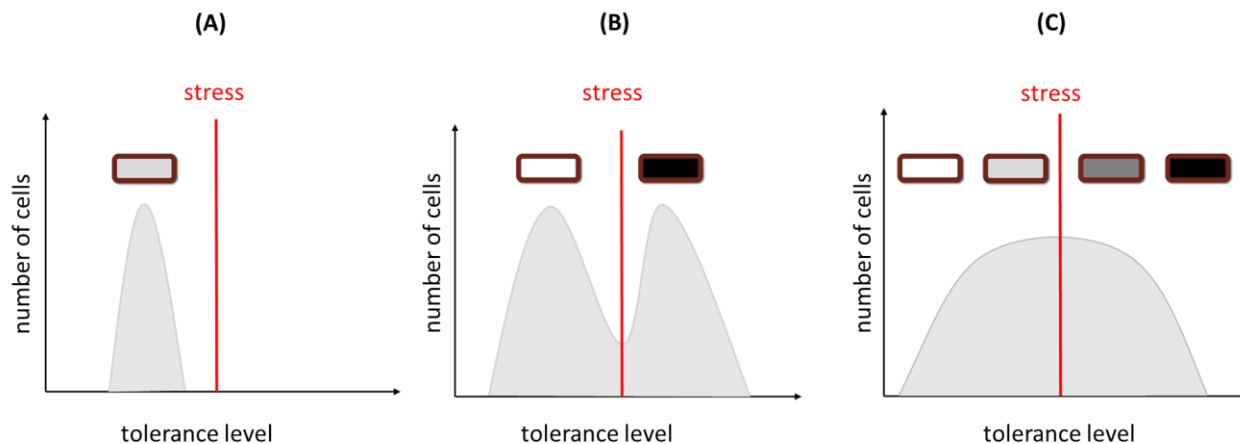


Figure 1.1 - Relationship between different types of heterogeneity and the outcome of exposure to a stress. A) If the stress level is beyond the tolerance level of all individuals or in cases where variations across individuals is not broad enough to overcome the stress, we observe all individuals as one phenotype. B) If the distribution of phenotypes has two peaks, like in the case of persistence, two phenotypes are observable. C) In case of broad variations of tolerance level across a population, we see different phenotypes survive the stress.

Bacterial persistence is one of the best-studied examples of discrete phenotypic heterogeneity. Bacterial persistence refers to a physiological state where, in an isogenic population, a rare subpopulation of cells enters dormancy. Dormancy allows these cells to escape the action of an antibiotic and thus leads to resistance (Bigger, 1944). Cancer cells have also been shown to have phenotypic resistance to drugs. Leukemic cells have variable resistance against vincristine (Pisco et al., 2013). Gene expression of the MDR1 protein, which is responsible for exporting the drug out of the cells, showed a continuous distribution across the population. This heterogeneity in gene expression leads to a continuous distribution of phenotypes.

In a simplified model, we can establish a relationship between the expression of genes and a phenotype in response to a stressor. In the simplest scenario, there is a one gene to one phenotype relationship and phenotypic variations across individuals could be mapped onto their gene expression profile (Figure 1.2, panel B). However, the picture is often much more complicated. Phenotype could be an outcome of the expression of multiple genes that encode subunits of a single protein or a number of different proteins expressed by a complicated regulatory network. To relate the expression of one gene (or a set of genes) to one tolerance phenotype, one must find gene(s) that contribute to the tolerance, quantify gene expression variation within a population and map them to the distribution of phenotypes.

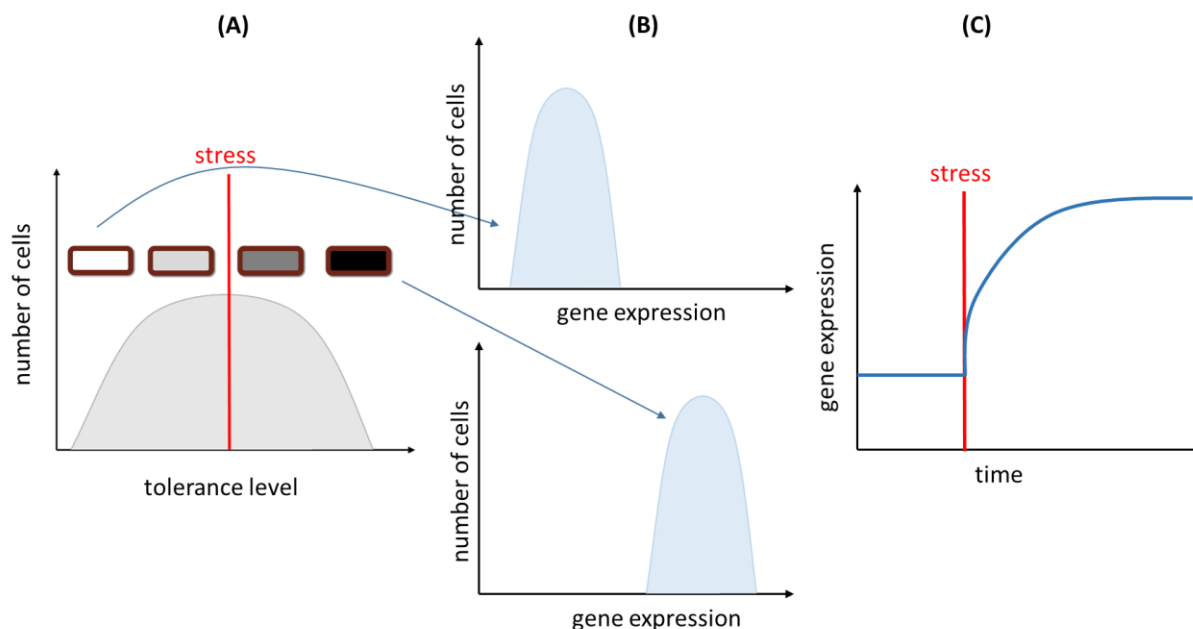


Figure 1.2 - Different gene expression profile between subpopulation is one of the possible mechanisms to observe heterogeneity. A) A broad continuous variation of phenotypes in facing a stress. B) Differences between phenotypes can emerge from different expression levels of a gene (or genes). C) Such genes may be up-regulated (or down-regulated) when facing the stressor.

Here we investigate the phenotypic heterogeneity and the gene expression profiles in response to a stressor in a model system where the stressor is generated intracellularly. *Methylobacterium extorquens* is a bacterium that naturally lives on plant leaves and consumes methanol secreted from the leaves. Methanol is oxidized to formaldehyde through the action of the methanol dehydrogenase (MDH) enzyme with a rate of 2 mM/s (Vorholt et al., 2000). At low concentrations, formaldehyde is non-toxic for the cell, but at higher concentrations it can be lethal.

In this study, we show that a population of *M. extorquens* exhibits a distribution of formaldehyde tolerant phenotypes. By tracking death and survival at different formaldehyde concentrations we demonstrate that, as in the case of resistance to cancer drugs, this distribution is continuous. Here, I use a PDE framework to model phenotypic heterogeneity to the toxicity of formaldehyde. With this model, I showed that formaldehyde-induced death is binary; there is an absolute threshold between survival and death, even though the distribution of tolerance to formaldehyde is continuous. In addition, this work showed that growth and selection by death are not able to explain the observed changes in phenotypic distributions through the timecourse of the experiment. Therefore, other mechanisms that enable cells to change their phenotypic state must exist. To understand the genes potentially involved in formaldehyde tolerance, we need to first identify genes that show changes in expression upon exposure to the stressor (Figure 1.2, panel C). Formaldehyde stress has two layers of

action. First, formaldehyde is a highly reactive agent with a potential to broadly damage proteins and DNA. Second, formaldehyde interacts with the EfgA protein, a formaldehyde sensor, and peptide-deformylase to stop translation (Nayak et al. in prep.). Therefore, the mechanisms of action of formaldehyde in *M. extorquens* are two-fold: it is comparable to general stress response inducers (like heat) and one-target translation inhibitors such as the antibiotic kanamycin. To characterize and differentiate the genes involved in each of the mechanisms of formaldehyde action (translation inhibition and general tolerance to the stress), I performed an extensive transcriptomic analysis of *M. extorquens* across an array of genotypes (wild-type and EfgA-deficient mutants), treatments (formaldehyde or kanamycin) and timepoints. I showed that almost half of the response between kanamycin and formaldehyde response is shared, thereby identifying common ground between the translational inhibitors and stark differences. In addition, I found that the EfgA specific response is a key factor in the total cellular response to formaldehyde. Further, I showed the response to formaldehyde involves some general stress response proteins such as chaperones, DNA repair system as well as cytochromes, ABC transporters and flagellar proteins. My analyses reflected the general and specific actions of formaldehyde, has begun to parse apart the cellular consequences of formaldehyde and EfgA, and generated a number of testable hypotheses.

As a student in Bioinformatics and Computational Biology program, this project has been a great opportunity for me to learn two different aspects of computational biology to explore two facets of the physiological response that *M. extorquens* has to a metabolic stressor. In the second chapter I did mathematical modeling to understand change in distribution of tolerance in face of the formaldehyde stress. In the third chapter I used a bioinformatics/statistics approach to analyze RNA-seq data and found genes involved in response to formaldehyde toxicity.

2 Mathematical Modeling of Response to Formaldehyde in *Methylobacterium extorquens*

2.1 Introduction

Stochastic processes play a vital role in a single cell. Many reactions depend on a small number of molecules, and are thus left sensitive to random fluctuations in their number. Depending upon how molecular details lead to higher-level phenotypes, this has the potential to generate significant “noise” in biological outcomes. Often, at the population scale this noise can be ignored, and the assumption that the whole population can be well described by the average behavior is sufficient. In other cases, the effect of noise can be extreme and can lead to two or more distinct phenotypes within the population, a phenomenon known as bistability or multistability.

Although, in principle, any differences between genetically identical cells are fairly described as phenotypic heterogeneity, the majority of studies of heterogeneity have been when there are large differences in phenotypes, and there is an assumed advantage for the genotype that can produce their phenotypes (Ackermann, 2015). Consider the examples of growth versus genetic resistance, or which substrate to utilize, where phenotypic heterogeneity has the potential to allow for survival in stressful conditions or allows cells to efficiently switch between nutrients. If a growing population at least occasionally produces cells that have increased resistance or distinct substrate use, even if these cells grow slowly or not at all, the heterogeneous genotype could survive a stressor or switch to a new substrate that a homogeneous one may not. This is the presumed ecological significance for competence in *Bacillus subtilis*. Stochastic variation in expression of ComK (Maamar et al., 2007) or phosphorylation of sporulation protein A (Spo0A) determines which cells become competent for DNA uptake (Chastanet et al., 2010), which in turn brings up the potential to use DNA as a nutrient and/or obtain an allelic variant that allows that genotype to survive. Similarly, yeast cells grown on a medium with low glucose and high galactose stochastically use either glucose or galactose. (Biggar and Crabtree, 2001).

One of the most famous examples of bistable heterogeneity in bacterial populations – and amongst the most relevant comparisons to the empirical system that I have modeled in this chapter – is bacterial persistence in the face of antibiotics. Bacterial persistence refers to a state where, in an isogenic population, a rare subpopulation of cells enters dormancy, and because of this dormancy they escape the action of an antibiotic (Bigger, 1944). Persister cells are thus physiologically different from the rest of the population despite having no genetic mutations (which would be termed “resistance”). This phenomenon often involves a toxin/anti toxin system (Korch et al., 2003), although other mechanisms

have been discovered (Van den Bergh et al., 2017). Mathematically, bistable persistence to antibiotics has been represented by a simple ODE model that describes the dynamics of sensitive cells and persisters (Figure 2.1) (Balaban et al., 2004).

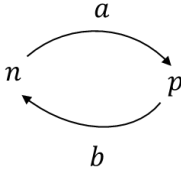
$$\begin{aligned}\frac{dn}{dt} &= -an + bp + \mu_n n \\ \frac{dp}{dt} &= an - bp + \mu_p n\end{aligned}$$


Figure 2.1 - Dynamics of persisters and sensitive cells (Balaban et al., 2004). Persisters are represented by p and natural (sensitive) cells by n . The per capita rate of switching the natural cells to persisters is denoted by a , and the per capita rate of reverse process is denoted by b . The per capita growth rates of natural cells and persisters are denoted by μ_n and μ_p , respectively.

Each of the cell types is assumed to grow at its own rate, and switch between cell types at two different rates. Once formulated in this manner, it is possible to explore the potential for a given strategy – a genotype with a particular set of parameters – to be selectively favored over another genotype, thereby illuminating the ecological regimes that could have selected for the emergence and maintenance of heterogeneity (Kussell et al., 2005). Although phenotypic heterogeneity for traits like persistence has been studied most extensively in bacteria, there are numerous eukaryotic examples, such as in *Candida albicans* (LaFleur et al., 2006) and *Saccharomyces cerevisiae* (Bojsen et al., 2016).

Although most examples of heterogeneity that have been studied involve discrete phenotypes, it is also possible that heterogeneity within a population is continuous (Chang et al., 2008; Pisco et al., 2013). Phenotypic resistance of HL60 leukemic cancer cells to anti-cancer drugs is one such example. MDR1, or multidrug resistance protein, is part of the ABC transporter protein family. Expression of this protein has a role in resistance to multiple drugs, which is known as the MDR phenotype in cancer cells (Gillet and Gottesman, 2010; Pisco et al., 2013). The expression pattern of MDR1 shows a continuous distribution, so phenotypic resistance to anti-cancer drugs could be expressed as a continuous trait (Pisco et al., 2013).

For phenotypic heterogeneity that is continuous, it is more appropriate to use a PDE model, rather than an ODE model between discrete categories (Lorenzi et al., 2016). In their model of leukemia, Lorenzi et al track the expression level of the drug resistance gene (such as MDR1) in each cell.

Numerous mechanisms have been proposed to generate significant phenotypic heterogeneity in populations (Ackermann, 2015). Perhaps the most common mechanism proposed is stochastic gene expression (Elowitz et al., 2002). A promoter in a single cell is often only rarely transcribed from,

producing a burst of transcripts, and a corresponding pulse of proteins. This generates variability in expression in a single cell with time, as well as between cells at any given moment of time. Simulation of gene expression events in a regulatory network showed that the time for accumulation of a regulatory protein is stochastic, and thus the level of gene expression within a population with the same genetic background is variable (McAdams and Arkin, 1997). In many cases, positive feedback loops exist that amplify the effect of stochastic gene expression, thereby leading to expression thresholds that, once crossed, generate discretely different cell fates (bistability or multistability) within a genetically homogeneous population (Acar et al., 2005; Hasty et al., 2000; Maamar and Dubnau, 2005; Maamar et al., 2007; Raj and van Oudenaarden, 2008; Smits et al., 2005; Süel et al., 2006; Süel et al., 2007; Weinberger et al., 2005). In addition to positive feedback loops, a double negative feedback loop can also lead to bistable or multistable behavior (Ferrell Jr, 2012; Plahte et al., 1995; Snoussi, 1998; Thomas, 1981). In addition to the effect of noise in gene expression on heterogeneity, a theoretical study shows that stochastic partitioning of macromolecules in cell division could mimic the effect of noise in gene expression (Huh and Paulsson, 2011). Asymmetric segregation of protein aggregates have been shown to play a role in ageing of cells and thus phenotypic variability within populations (Lindner et al., 2008). Pole age of a cell also has an effect on its size and division time. As an example, in *Methylobacterium extorquens* AM1, cells size increase and division time decrease with increasing pole age (Bergmiller and Ackermann, 2011).

In this project, I have investigated phenotypic heterogeneity in response to an internal metabolic toxin. Like the case of cancer, and unlike in bacterial persistence, this heterogeneity is continuous rather than discrete. In this chapter, I use dynamic mathematical models in conjunction with experimental data to determine how cell death depends on the formaldehyde concentration and the tolerance state of cells. Furthermore, I also demonstrated that growth and selection are not sufficient to explain the experimental data, and there appears to be different active mechanisms changing the tolerance phenotypes under different environmental conditions. The models provide support that there are mechanisms such as gene expression or other regulatory events that could play a role in phenotypic changes.

2.2 Empirical model system and background data

Our model system where we have discovered a novel type of phenotypic heterogeneity involves *Methylobacterium extorquens* PA1, and its response to varying concentrations of a toxin, formaldehyde. *M. extorquens* is an aerobic, facultatively methylotrophic Alphaproteobacterium and lives on plants as an epiphyte (Vorholt, 2012). It has been the premier model system to study single-C (C_1) metabolism for over 60 years. *M. extorquens* can grow upon C_1 substrates, such as methanol and

formaldehyde, as well as some multi-carbon substrates, such as succinate. Methanol enters the periplasm and it is oxidized by methanol dehydrogenase to formaldehyde. The central carbon and energy metabolic pathways of *M. extorquens* are responsible for the oxidation of formaldehyde in the cytoplasm to formate, and for the subsequent fork that allocates formate either to further oxidation to carbon dioxide, or incorporation into biomass (Crowther et al., 2008; Marx et al., 2003; Marx et al., 2005).

This project began with an experimental quantitative analysis by Dr. Jessica Lee of how formaldehyde affects the population dynamics of *M. extorquens*. An inoculum of *M. extorquens* PA1 – strain CM2730, which is wild-type other than a deletion of the cellulose production locus to prevent clumping; (Delaney et al., 2013b) – was taken from a stationary-phase culture, and then cultured in media with methanol (15 mM) and different concentrations of externally added formaldehyde. Viability was tracked via cfu/ml on plates with just methanol. At low concentrations of formaldehyde (<2 mM) no death was observed, whereas at a concentration of 4 mM there was a decline in cell viability before growth. At levels of formaldehyde higher than this, the viable cell counts showed a log-linear relationship with time, with a slope that increased with elevated concentrations of formaldehyde (Figure 2.2). The simple exponential decay of viability matches the kinetics seen for antibiotics (e.g. Udekwu et al., 2009).

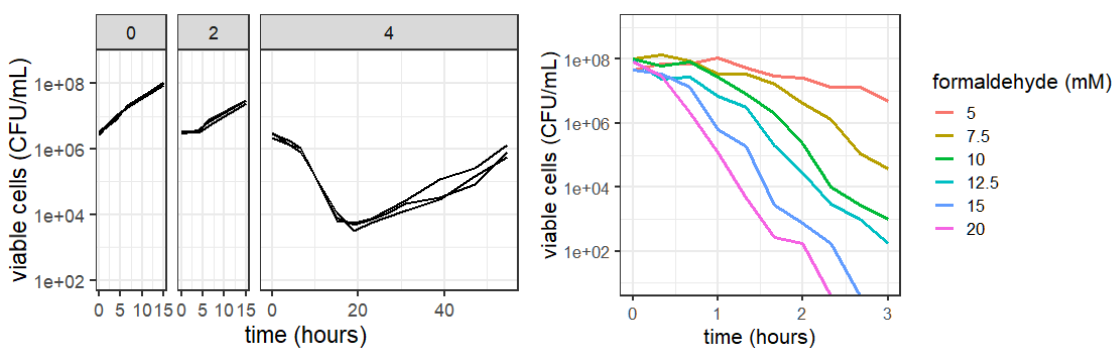


Figure 2.2 - Dynamics of cells grown on different concentrations of formaldehyde. Left: growing on low to medium concentrations of formaldehyde. Right: death dynamics on high concentrations of formaldehyde.

Four hypotheses were proposed to explain this phenomenon. First, cells may have consumed formaldehyde as they die, and when the concentration reached a tolerable level, they regrew. This kind of phenomenon has been observed for killing with β -lactams, due to the release of β -lactamase into the media leading to a decreased level of stressor (Artemova et al., 2015). Second, the regrowth could represent cells bearing a genetic mutation making them resistant to this level of formaldehyde. Third, cells may have lost the ability to make colonies, but were not completely inviable, a state known in

environmental microbiology as viable but not culturable VBNC (Pinto et al., 2015). The subsequent increase in cfu/ml could have been due to the return of cells out of the VBNC state to a growing state. Fourth, there could have been phenotypic heterogeneity in the population for the ability to grow in the presence of formaldehyde. Under this scenario, most of the cells would have died during the first 20 hours, but a minority of cells could have already been present that were growing during this time, ultimately taking over the population.

The first experiment to distinguish between these possibilities was to track the formaldehyde concentration in growth on methanol treated with 4 mM formaldehyde (Figure 2.3). This revealed that there was no change in formaldehyde concentration until ~70 h, long after the change in the trajectory of cells. This ruled out the first hypothesis that the environment of the cells had improved.

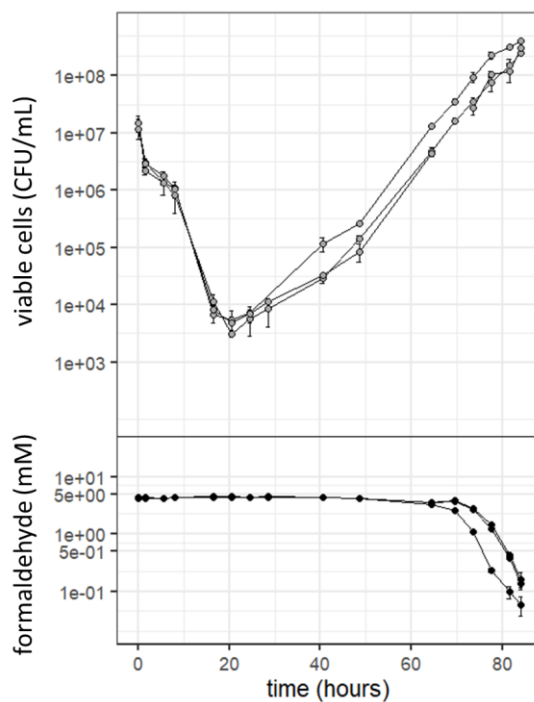


Figure 2.3 - Viability and measured formaldehyde data for growth on 4 mM formaldehyde. Top: viability of cells show an initial decline and then regrowth. Bottom: measured formaldehyde shows no change in concentration during the course of growth.

Next, gDNA from the population at the end of 80 hours was prepared and submitted for whole genome sequencing. This revealed that no mutations or gene amplification (e.g., (Reams et al., 2012)) occurred over this timespan, ruling out the second hypothesis. Furthermore, the ability of the culture to grow on 4 mM formaldehyde was retested. Using cells taken immediately from 80 hours, these cells could grow immediately in the presence of 4 mM formaldehyde. In contrast, if the cells were inoculated into the

succinate medium for a full growth cycle and then retested, this ability was lost. These data ruled out mutations as an explanation, suggesting that some type of phenotypic change was occurring.

Two complementary experiments were performed to determine whether all cells slowly recovered (hypothesis 3), or if a subpopulation was already growing throughout the experiment (hypothesis 4). First, to track the proliferation of cells in liquid media, the culture was treated with fluorescent linker dye that stains membranes, and trajectories of cell number and per cell fluorescence were observed by flow cytometry. In each generation, the number of cells would double and every cell would become half as labeled due to dilution by newly synthesized membrane components. When cells were grown on methanol with no formaldehyde (Figure 2.4, 0 mM panel), there was a smooth, unimodal increase in cell number and decrease in fluorescence. In contrast, when the culture grew with 4 mM formaldehyde, the distribution of cells initially neither increased in number nor exhibited a decrease in per cell fluorescent membrane dye. By 37 hours there was, however, a clear subpopulation that emerged that was already much less bright than the initial distribution. This subpopulation increased in number until it ultimately dominated the distribution.

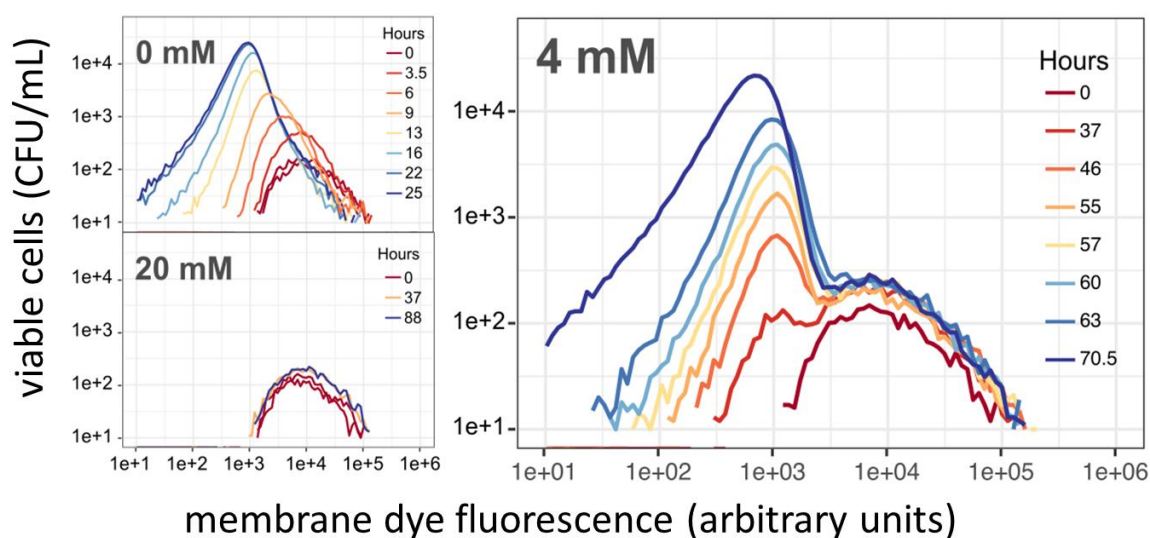


Figure 2.4 - Fluorescent membrane dye showed presence of a subpopulation in growth on formaldehyde. Cell proliferation assay showed that a *M. extorquens* population exposed to moderate levels of formaldehyde (4 mM) has both growing and non-growing subpopulations. Cells were stained with fluorescent membrane dye then allowed to grow in the presence of 0, 4, or 20 mM formaldehyde. Histograms show per-cell fluorescence of the cells present at each time point. Top, left: media with no formaldehyde, all cells undergo doubling, diluting their membrane. Bottom, left: at high concentrations of formaldehyde, no cells grow, leaving per-cell fluorescence unchanged. Right: in the presence of 4 mM formaldehyde, most cells do not grow, but a few do; so a small population with lower per-cell fluorescence becomes detectable at 37 hours and grows afterward.

The second approach was to follow the growth of single cells using time-lapse video microscopy of populations grown on agar pads. This work, performed by Shahla Nemati in the laboratory of Dr. Andreas Vasdekis, allowed the morphology and division time of individual cells and their progenies to be observed. Cells were treated with either 0 or 2.5 mM of formaldehyde (and 15 mM methanol) and trajectories were observed. Nearly every cell grown on methanol without formaldehyde formed a colony, with relatively little spread in times of initiation. In contrast, in the population that experienced 2.5 mM, only 1.97% were able to grow at all (Figure 2.5). For those cells that did grow in the presence of 2.5 mM formaldehyde, their doubling times were indistinguishable from the cells grown without formaldehyde. Remarkably, no partial growth phenotype was observed. The majority of cells that did not form a colony did not elongate, nor show any detectable change in cell morphology.

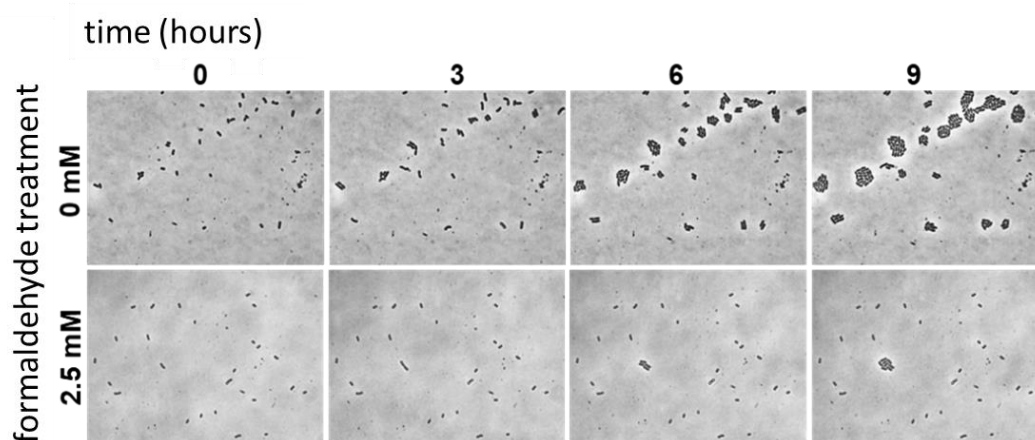


Figure 2.5 - Time-lapse microscopy showed bimodal (i.e., growth or non-growth) phenotypes in response to formaldehyde. Cells were embedded in agar medium with methanol and either 0 mM (top) or 2.5 mM (bottom) formaldehyde and monitored for 9 hours (~3 generations). At 0 mM, 256 cells were observed and all underwent at least one doubling; at 2.5 mM, 546 cells were observed and 11 (1.97%) underwent at least one doubling.

These experiments provide solid evidence for phenotypic heterogeneity in the ability to grow in the presence of formaldehyde at levels sufficiently high to cause the majority to die. We call this phenomenon phenotypic heterogeneity of tolerance, and emphasize that it is distinct from persistence, for the latter describes resistance due to being in a non-growing phenotype, whereas the tolerant cells that we have observed grow equivalently to non-tolerant cells in the absence of the stressor.

Is formaldehyde tolerance in *M. extorquens* a discrete trait, or continuous? In order to determine the shape of the distribution of formaldehyde tolerance, cells were plated onto agar plates with methanol (15 mM) and different concentrations of formaldehyde between 1 and 10 mM, at 1 mM increments. Observing a colony at a given level of formaldehyde reveals that a subset of cells is tolerant to at least

that level of formaldehyde. As such, the measured distribution is cumulative. If the distribution were discrete, two (or more) plateaus would be observed across the range of concentrations (Figure 2.6, left). In contrast, the data obtained reveal tolerance is continuous, unimodal and approximately exponentially decreasing (Figure 2.6, right). This distribution has been assayed many times, and the overall shape is highly consistent for a particular condition, but does show subtle differences between pre-growth conditions (i.e., methanol vs. succinate, or exponential vs. stationary-phase cells; data not shown).

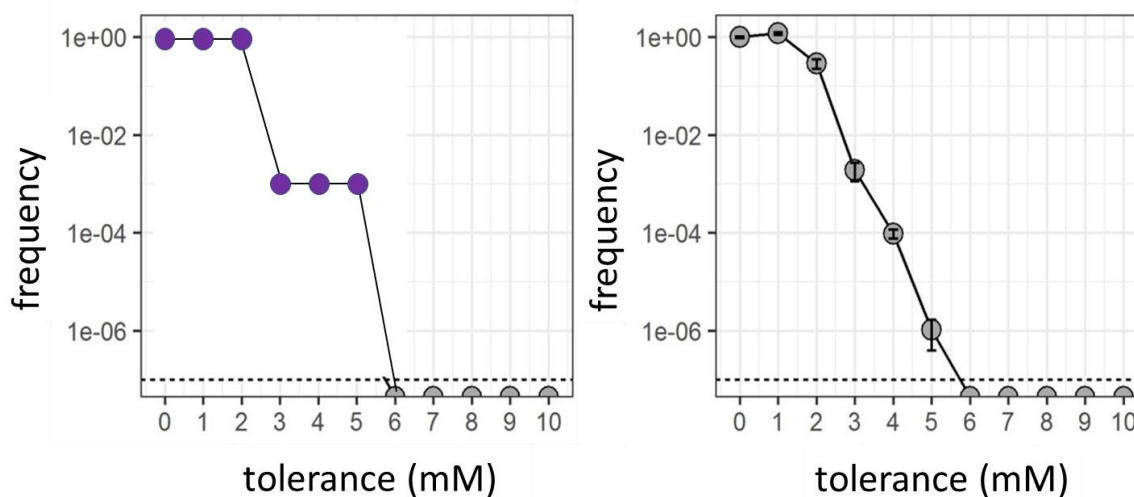


Figure 2.6 - Continuous distribution of tolerance to formaldehyde. Left: in the case of a bimodal discrete distribution, two distinct peaks would be observed. Right: observed distribution on agar plates containing different concentrations of formaldehyde show a cumulative, continuous and exponentially decreasing distribution, confirming tolerance to formaldehyde is continuous rather than discrete.

Returning to the above experiment of net death transitioning to growth with 4 mM formaldehyde, the full spectrum of tolerances was now assessed, rather than just cells that could grow with 0 or 4 mM formaldehyde (Figure 2.7). All subpopulations with tolerance level below 4 mM decreased in frequency, whereas those above 4 mM increased in frequency. Furthermore, cells with tolerance between 5 to 8 mM became detectable over this timescale. Thus, the tolerance distribution is not only continuous but also changes shape over the time scale of this experiment.

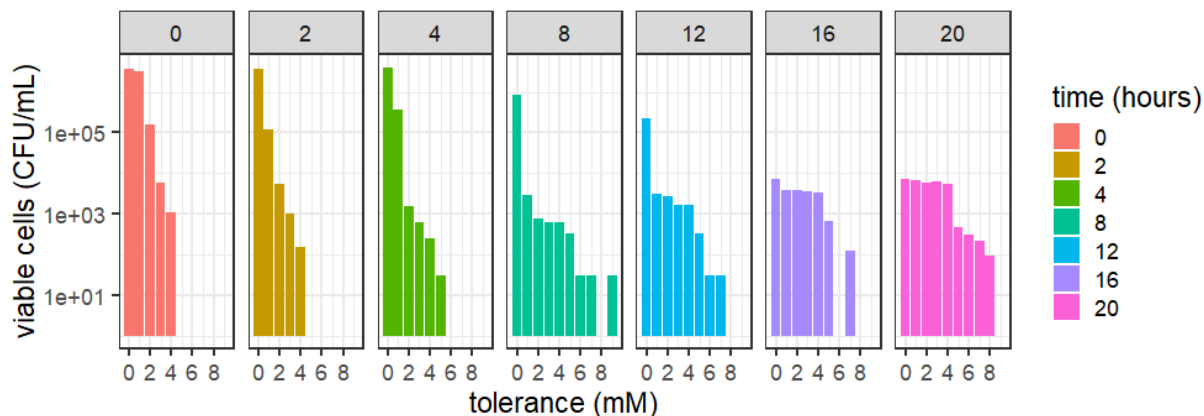


Figure 2.7 - Dynamic of tolerance distribution in growth on methanol with 4 mM formaldehyde. Subpopulations with tolerance levels below 4 mM decrease in frequency and those with higher tolerance levels increase frequency. The data are cumulative; each tolerance level gives the number of cells that are able to tolerate at least that concentration of formaldehyde.

Once the population has achieved high tolerance due to exposure to 4 mM formaldehyde, what occurs during the loss of tolerance? An inoculum from the culture at 96 hours was cultured in a medium with either methanol or succinate, but no formaldehyde. In the succinate medium, cells in higher tolerance levels were rapidly lost, and within 24 hours (~6 doublings) the population returned close to the original distribution of tolerance. In contrast, same experiment in the methanol medium revealed that tolerance was maintained in all subpopulations (Figure 2.8). Maintaining tolerance in regrowth of the selected population on methanol shows that there is hysteresis, where the behavior of a population depends on its previous condition (Deris et al., 2013; Igoshin et al., 2008; Savageau, 1999). These changes in tolerance could be due to either regulatory or epigenetic changes that shift tolerance levels, and/or from demographic changes in the population due to differential death and growth of cells at different tolerance levels.

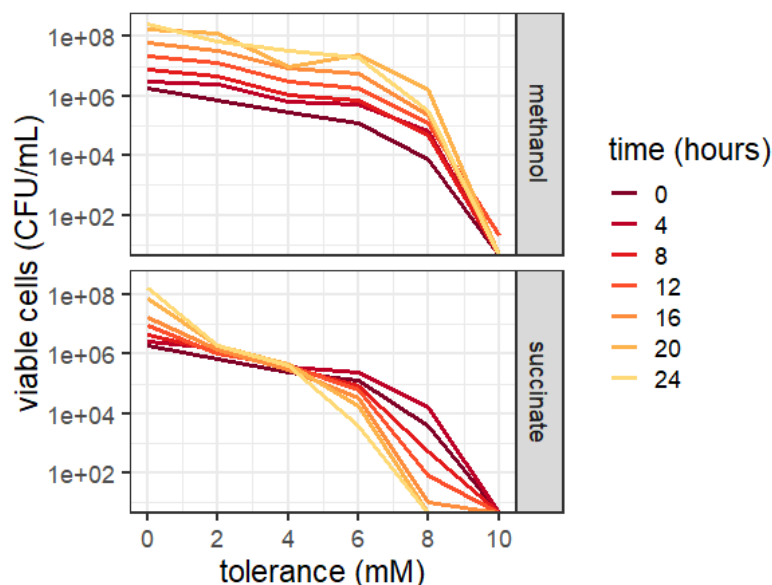


Figure 2.8 - Change in tolerance distribution in regrowth on methanol or succinate. The top row shows that a selected population grown with formaldehyde preserves its tolerance level in the methanol media whereas in succinate media (bottom row) the tolerant subpopulations decline. The data are cumulative, such that each tolerance level shows the cells that are able to tolerate at least the given concentration of formaldehyde.

2.3 Model development

We have already seen that exposing a population of *M. extorquens* to high concentrations of formaldehyde caused death, with an increasing exponential decay rate as the concentration of formaldehyde increased. There are subpopulations of bacteria with different tolerance levels to toxicity of formaldehyde; tracking the growth in a culture treated with formaldehyde showed a change in the distribution of tolerance. In addition, re-growing an already selected population in formaldehyde had different consequences in different media. In a methanol environment the cells kept their tolerance to formaldehyde, whereas in succinate, cells with high tolerance levels lost their tolerance. I have developed a mathematical model to determine how growth, death, and changes in tolerance of individual cells affect the distribution of phenotypic tolerance in a population. Using this mathematical model, I sought to address two basic questions. First, what is the relationship between the concentration of formaldehyde, the phenotypic state of cells, and death rate? Second, is there evidence for movement between phenotypic states on the time-scale of any of these experiments? (Figure 2.9)

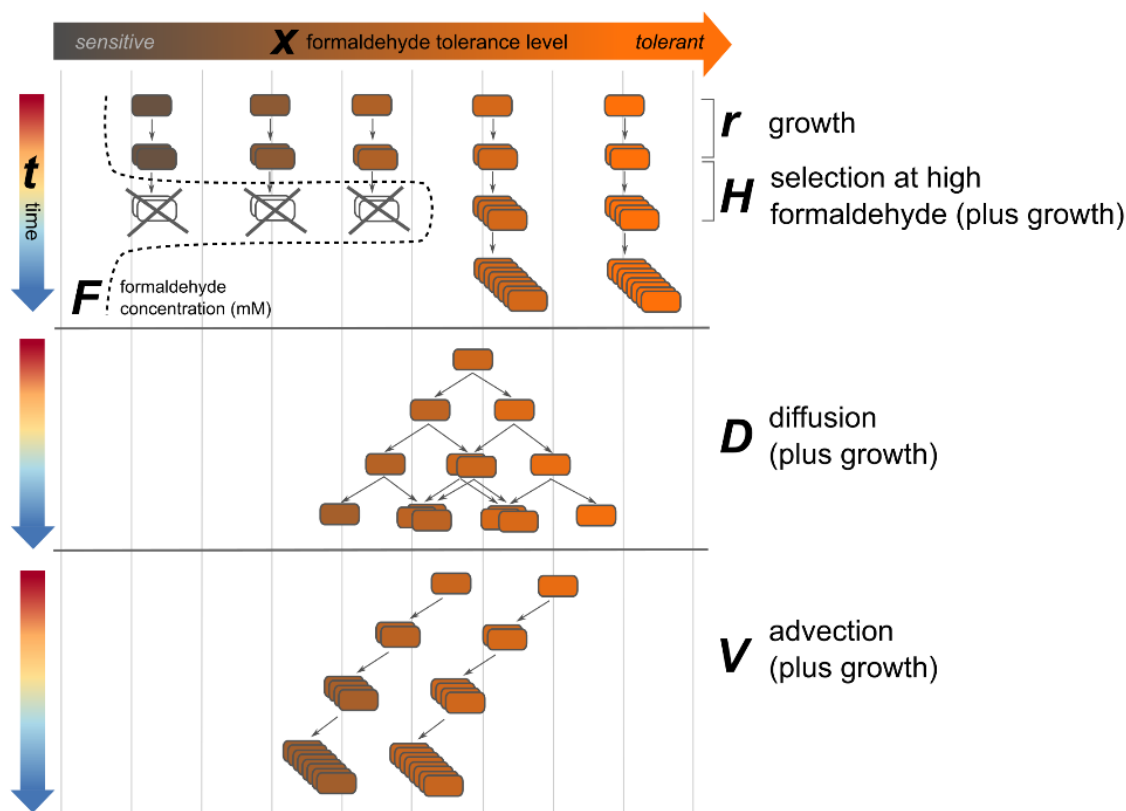


Figure 2.9 - Schematic diagram of growth and phenotypic changes over time (from Dr. Jessica Lee). The vertical axis is the time variable in the model. The horizontal axis is the spatial variable x in the model that shows the tolerance level. Formaldehyde concentration is shown by dashed line in the top panel. In a spike of formaldehyde, cells at the lower concentrations of formaldehyde are not able to survive and only those having higher tolerance levels keep growing. When the formaldehyde is gone there are two processes that change phenotypic state of cells: diffusion move cells bi-directionally in phenotypic space, and advection brings cells to the lower tolerance levels.

The model tracks concentration of cells (N), methanol (M), succinate (S) and formaldehyde (F) as they are utilized by growing cells with Michaelis-Menten kinetics in a well-mixed homogeneous environment. To capture the continuous phenotypic tolerance, I developed a system of partial differential equations (PDEs) with a time variable and a continuous, 'spatial' variable x corresponding to a cell's phenotypic state. The variable x represents the highest concentration of formaldehyde in which a cell can grow without death. Boundary conditions for $N(x, t)$ are given by: $\frac{\partial N(0, t)}{\partial x} = 0$ and $\frac{\partial N(L, t)}{\partial x} = 0$. The upper boundary (L) was set at $L = 10$ as no cell was observed in any experiment with tolerance level higher than 10 mM. Cells die at a rate that depends upon their phenotype x and the formaldehyde concentration. In contrast, the fact that after an initial decline, the population of cells

grows at or near typical rates argues that the growth rate does not need to be expressed as a function of either F or x . These assumptions lead to the following mathematical model:

$$\frac{dS(t)}{dt} = -V_{maxs} \left(\frac{S(t)}{S(t) + K_s} \right) \int_0^L N(x, t) dx \quad (1)$$

$$\frac{dM(t)}{dt} = -V_{maxm} \left(\frac{M(t)}{M(t) + K_m} \right) \int_0^L N(x, t) dx \quad (2)$$

$$\frac{dF(t)}{dt} = -V_{maxf} \left(\frac{F(t)}{F(t) + K_f} \right) \int_0^L N(x, t) dx \quad (3)$$

$$\frac{\partial N(x, t)}{\partial t} = r_m \frac{M(t)}{M(t) + K_m} N(x, t) - H(x, F) N(x, t) + D \frac{\partial^2 N(x, t)}{\partial x^2} + v \frac{\partial N(x, t)}{\partial x} \quad (4)$$

State variables and their units are shown in Table 2.1 and parameters in the model are given in Table 2.3. Equations 1-3 show changes in concentrations of the substrates: succinate, methanol and formaldehyde. Equation 4 shows change in the number of cells when they grow on methanol; in the case of growth on succinate, M could be replaced by S and r_m by r_s . The function $H(x, F)$ describes the per capita death rate as a function of tolerance level and concentration of formaldehyde.

Table 2.1 - Description of state variables and their units in the model.

State variables	Description	Units
S	Concentration of succinate	mM
M	Concentrations of methanol	mM
F	Concentration of formaldehyde	mM
N	Number of cells per ml of liquid culture	$cell\ ml^{-1}$

To investigate the relationship between concentration of formaldehyde and death rate, I studied two extreme hypotheses for how the per capita death rate depends upon tolerance. The first hypothesis is that the death rate can be expressed as an absolute threshold, whereby all cells with a value of x less than the external formaldehyde concentration die at an equal rate, whereas those equal to or greater

than this threshold do not die. The second hypothesis is that death can be expressed as gradually increasing the farther a cell's tolerance x is below the external formaldehyde concentration. Again, cells with values of x above the formaldehyde concentration do not die (Figure 2.10).

$$H_1(x, F) = \begin{cases} \propto F & \text{if } (x < F) \\ 0 & \text{if } (x \geq F) \end{cases} \quad (5)$$

$$H_2(x, F) = \begin{cases} \propto (F - x) & \text{if } (x < F) \\ 0 & \text{if } (x \geq F) \end{cases} \quad (6)$$

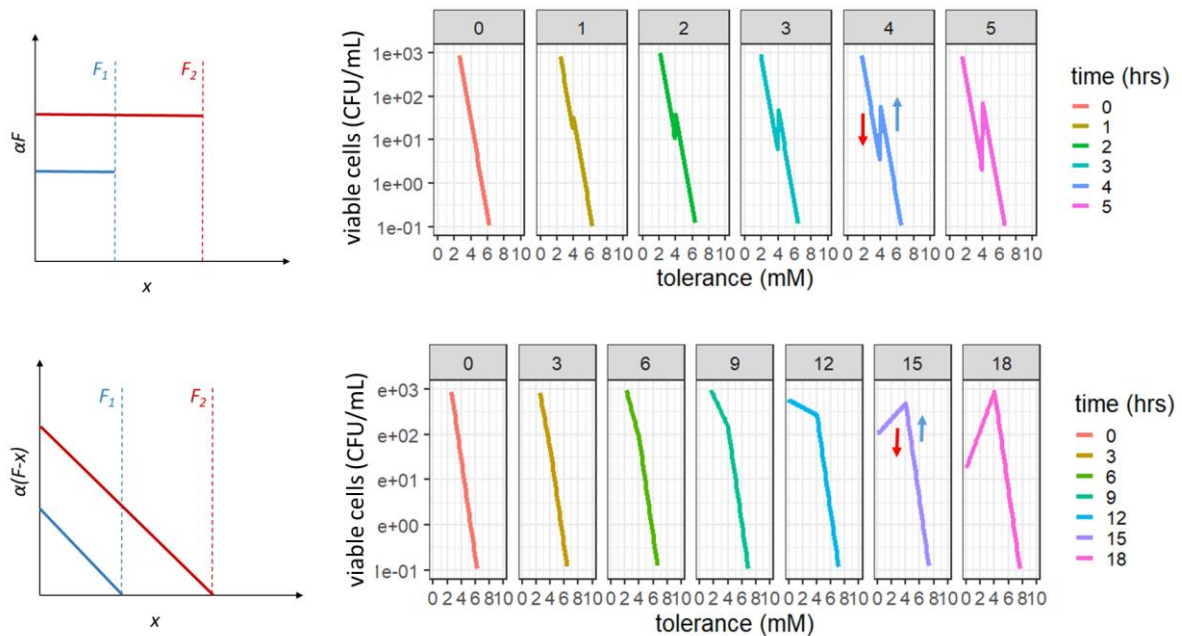


Figure 2.10 - Death function and how it affects the population's distribution. Left column: the horizontal axis shows the tolerance level of cells, and the vertical axis shows the death rate. F_1 and F_2 represent two different concentrations of formaldehyde. Right column: simulation of two different death functions and how they affect an exponential distribution of cells. Top row: death could be expressed as an absolute rate given the concentration of formaldehyde (Equation 5). Bottom row: the alternative hypothesis expresses death rate as a difference between a tolerance state and a concentration of formaldehyde (Equation 6).

To investigate how changes in phenotypic state affect the population dynamics, I included two types of movement in phenotypic space: diffusion and advection. Random, bidirectional changes in the phenotypic state of cells can be expressed with a diffusion operator with D as the diffusion coefficient. Diffusion spreads cells in both directions (gaining and losing tolerance) in phenotypic space.

Advection with advection coefficient v moves cells' tolerance in a single direction; a positive v leads to lower tolerance while a negative v increases tolerance.

To simplify the model, I make further assumptions regarding the methanol and formaldehyde concentration during the growth phase. Data on consumption of formaldehyde shows that the formaldehyde concentration is effectively unchanged until ~ 70 h, whereas the critical change from net death to growth occurs at 20 h. As the K_m for methanol and succinate is small, the methanol dehydrogenase (MDH) enzyme and succinate transporter are always effectively saturated during growth. With all of the external concentrations effectively unchanged during the key phenomena I seek to address, I set S , F and M to be constant in Equations 1, 2 and 3 yielding Equation 7. Throughout the rest of the chapter, all results correspond to the simplified model, Equation 7.

$$\frac{\partial N(x, t)}{\partial t} = r_m N(x, t) - H(x, F)N(x, t) + D \frac{\partial^2 N(x, t)}{\partial x^2} + v \frac{\partial N(x, t)}{\partial x} \quad (7)$$

2.4 Methods

2.4.1 Converting cumulative data to densities

As previously discussed, tolerance distributions were generated by plating *M. extorquens* on different concentrations of formaldehyde. Each category gives the number of cells with a given maximum tolerance level as observed by plating. These categories are cumulative in the sense that each contains the number of cells with that tolerance level and above (i.e., a cell that survives at 4 mM may have been able to survive a higher concentration). To calculate the non-cumulative densities at each category, differences between the bins in tolerance distribution were calculated as:

$$N(x) = \widehat{N}(x) - \widehat{N}(x + h) \quad (8)$$

where $N(x)$ is the number of cells in calculated densities (i.e., the number of cells that uniquely have a given tolerance level, (x)). $\widehat{N}(x)$ is the number of cells in the experimental cumulative distribution, and h is the step size between the categories in the data.

Experimental cumulative densities (Figure 2.7 and Figure 2.8) were converted to non-cumulative densities (Figure 2.11) using Equation 8. When grown on moderate concentrations of formaldehyde and methanol media, cells transiently lose tolerance from time 2 hours to time 0 and then progressively gain tolerance (Figure 2.11, top). The mechanism of this small transient shift is unclear for now, and thus we are not modeling this part of the phenomena. Because of this, all of the initial distributions in the model refer to time 2 (hours).

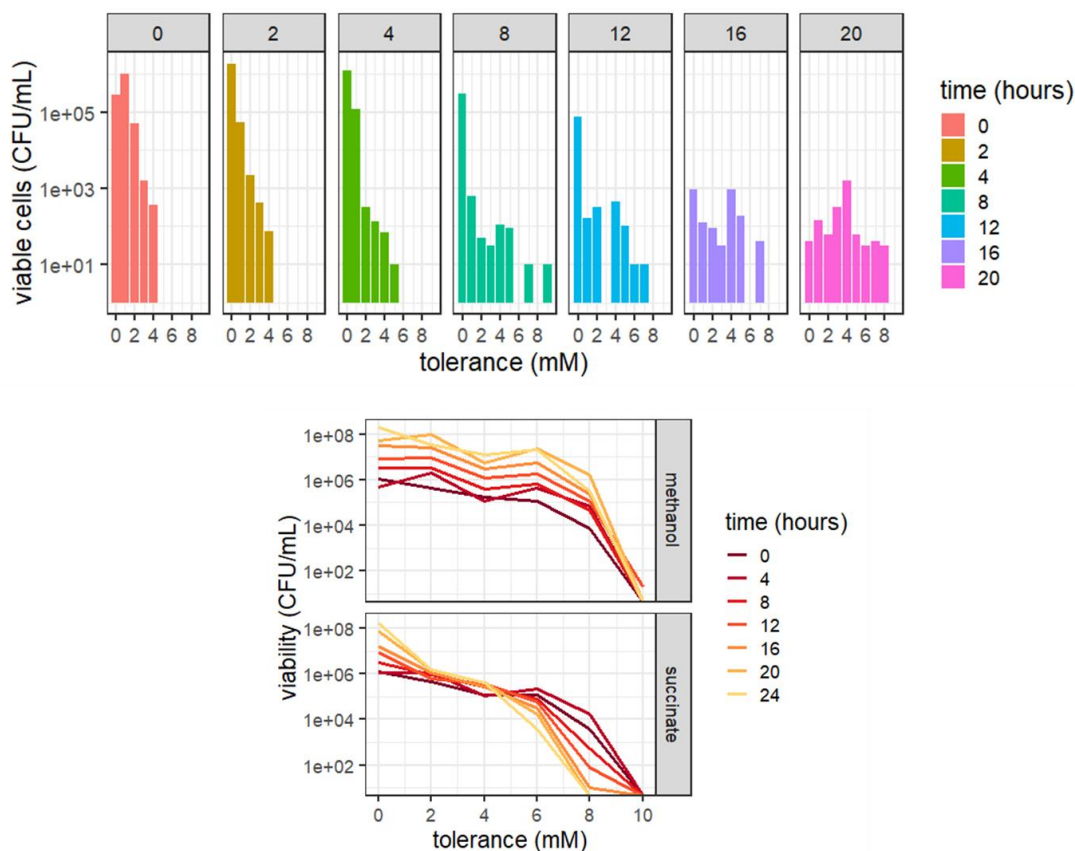


Figure 2.11 - Non-cumulative densities calculated from taking differences of cumulative densities. Top: densities calculated from tolerance change in growth on methanol with 4 mM formaldehyde. Bottom: calculated densities from the regrowth on methanol or succinate data with no formaldehyde.

2.4.2 Initial conditions

The initial condition for the number of cells with a given tolerance comes from the measured data at time 2 hours; (see previous section) however, the limit of detection must be considered. In the experiment each spot of liquid culture in the plate is 10 μl . There were 3 spots of per dilution per sample. The lowest detectable abundance of cells per sample would be 1 cell in 30 μl , which is approximately 33 cells ml^{-1} . I investigated the effect of extending the initial distribution below the limit of detection using an exponential distribution (Figure 2.12). The minimum allowed number in the exponential fit was set to 0.05 cells ml^{-1} ; in the model the volume of the culture is 1 ml and in the experiment volume of a flask is 20 ml, so the minimum possible number in extending the distribution is 1 cell in 20 ml which equals 0.05 cells ml^{-1} in the model. As the model tracks a continuous tolerance distribution, I interpolated the measured initial tolerance data using cubic interpolation (i.e., low degree polynomials in each interval of interpolation to smoothly interpolate the data) with the `splinefun()`

function in R. This results in an initial distribution with a finer resolution set to 0.01 mM differences in concentration, resulting in 1001 bins rather than the 11 that were measured (Figure 2.13).

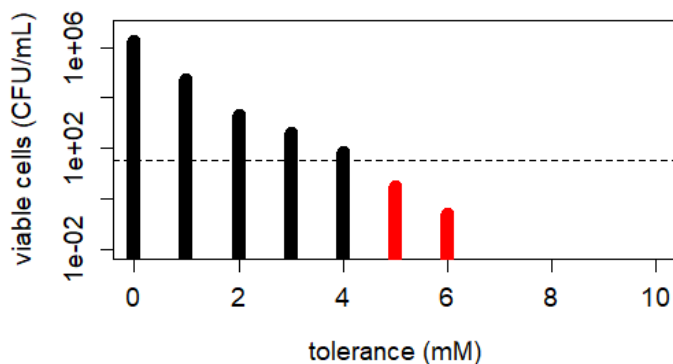


Figure 2.12 - Extending the initial condition beyond the limit of detection. As there is a limit of detection in the observations, the actual distribution likely has bins that are below the level of detection (horizontal dashed line showing 33 cells). The red bins were added using an exponential fit of the data.

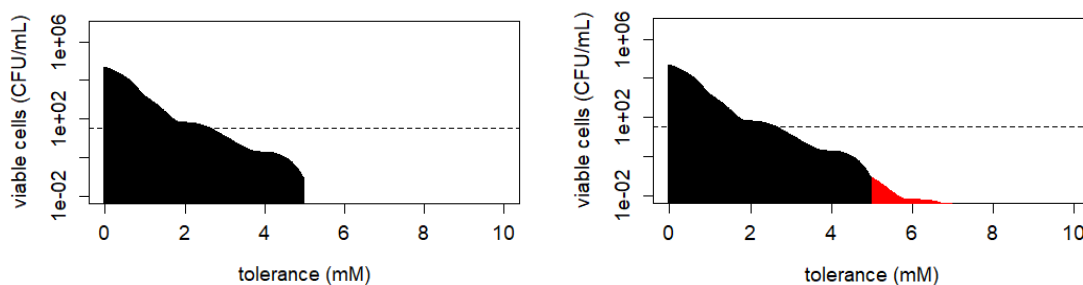


Figure 2.13 - Original distribution and extended distribution. Left: continuous distribution generated from the discrete data using cubic interpolation. Right: continuous distribution generated from discrete data with added bins from the exponential fit.

2.4.3 Parameter estimation

Growth rates and standard errors were estimated by fitting the exponential part of the growth curves on either succinate or methanol media. The `lmer()` function in the `lme4` library (Bates et al., 2014) was used to account for 3 replicates of each data point. Each data point is a cell number in log scale as the response variable, and time as the explanatory variable. The V_{max} values were calculated to set the time for consumption of substrates. The diffusion and advection parameters were estimated using tolerance distribution data over time (see supplementary for details on parameter estimation). The optimization algorithm for estimating the advection and diffusion parameters maximizes the log likelihood under a

linear model using the `lm()` function in R. Under this model I assumed the error is distributed normally between the data and model:

$$P(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

The function $P(x)$ gives the likelihood of the model value x given data μ , and the parameter σ is the standard deviation. The log likelihood can be written as:

$$LP(x) = -\ln(\sigma) - \ln\sqrt{2\pi} - \frac{(x - \mu)^2}{2\sigma^2}$$

The term $\ln\sqrt{2\pi}$ is constant, so for maximization the log likelihood can be rewritten as:

$$LL = -\ln(\sigma) - \frac{(x - \mu)^2}{2\sigma^2}$$

The above log likelihood is for one observation. The best-fit model values \hat{x}_i are those that maximize the log likelihood of all of the observed data:

$$\hat{x}_i = \operatorname{argmax}_{x_i} \left(\sum_{i=1}^N LL(x_i | \mu_i, \sigma_i) \right)$$

Here x_i is the model value and μ_i is the data value. The standard deviation σ_i can be calculated with the following equation:

$$\sigma_i = S_y \sqrt{\frac{1}{n} + \frac{(x_i - \bar{x})^2}{(n-1)S_x^2}}$$

where S_y is the standard error of the data, n is the number of categories in data, x_i is the model value, \bar{x} is the mean of all model values and S_x^2 is the standard error of all model values.

The hyperbolic arcsin (`asinh`) transformation was used to reduce skew; this transformation is similar to the log transformation, but can accommodate zero values in the data (Johnson, 1949):

$$\operatorname{asinh}(y) = \log(y + (y^2 + 1)^{1/2}).$$

The log likelihood function was maximized in R using the `optim()` function in R with the ‘‘Nelder-Mead’’ method (Nelder and Mead, 1965). Standard errors for the advection and diffusion parameters were calculated using the Hessian matrix obtained from the numerical optimization. The Hessian

matrix contains the second order partial derivatives of the log likelihood function evaluated at the maximum likelihood estimate. See Table 2.3 for parameters estimates.

2.4.4 Death functions

To investigate the relationship between the concentration of formaldehyde and viability, populations of *M. extorquens* were cultured in methanol liquid media with different concentrations of formaldehyde and viability was tracked on agar plates with no formaldehyde. The death curves show an exponential decline of viability with increasing concentrations of formaldehyde (Figure 2.14).

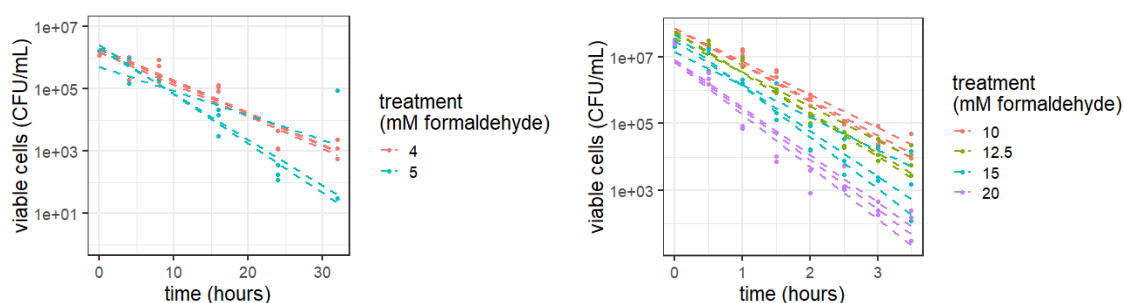


Figure 2.14 - Death curves show an exponential decline in viability with increasing concentrations of formaldehyde. Left: viability over time in low concentrations of formaldehyde, dots are the data points and dashed lines are the fits. Right: viability in high concentrations of formaldehyde.

As previously mentioned, I modeled death in two ways (Equations 5 and 6). The first way I modeled death assumes that the death rate of all cells with tolerance below the formaldehyde concentration is proportional to the formaldehyde concentration, and does not depend on tolerance (Equation 5). Death rates in Figure 2.14 were fit to Equation 5 using linear regression with the `lm()` function in R. I assumed that the death rate of cells when there is no formaldehyde ($F = 0$) is zero (Figure 2.15). This resulted in an estimated $\alpha = 0.189 \pm 0.010$.

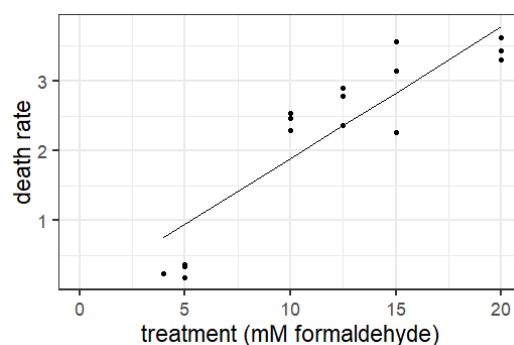


Figure 2.15 - Death rates fitted using the absolute death function. Death rate shows a linear relationship with growing concentration of formaldehyde.

The second way I modeled death assumes that the death rate of cells depends on both the formaldehyde concentration and the cells' tolerance level (Equation 6). For fitting this equation to death rates, I calculated the mean x value of cells from the initial tolerance measurement data. This value was obtained using a weighted mean of the continuous distribution (Figure 2.13). Specifically, $\bar{x} = \int_0^{\infty} f(x)x dx$, where $f(x)$ is the frequency of each tolerance level and x is the tolerance level. This yielded an estimate $\bar{x} \approx 0.271$ mM. Using linear regression with the `lm()` function in R, I estimated α as: 0.193 ± 0.010 .

2.4.5 Numerical solution of the PDE

The PDE was solved numerically by vectorized ODEs, where each ODE corresponded to a discrete bin within tolerance level of 0.01 mM formaldehyde. A finite difference grid was created using the `setup.grid.1D()` function, and advection and diffusion were calculated using the `tran.1D()` function, in the `Reactran` package v. 1.4.3.1 for R (Soetaert and Meysman, 2012). The vectorized ODEs were solved using the `ode.1D()` function from the package `deSolve` (Soetaert et al., 2010) in R, with the `lsoda` method (Hindmarsh, 1983).

2.4.6 Aggregating results of the continuous model to discrete bins

The numerical results of the PDE give continuous distributions, but the experimental data are only measured at discrete tolerance values. To compare the data with the model output, I aggregated the modeled tolerance distribution of cells to 10 bins (Figure 2.16 and Figure 2.17). For example, the model results corresponding to the subpopulations with tolerance levels of 0 mM to 0.99 mM would be aggregated into the 0 mM bin, and the subpopulations with tolerance levels from 1 mM to 1.99 mM would be aggregated to the 1 mM bin.

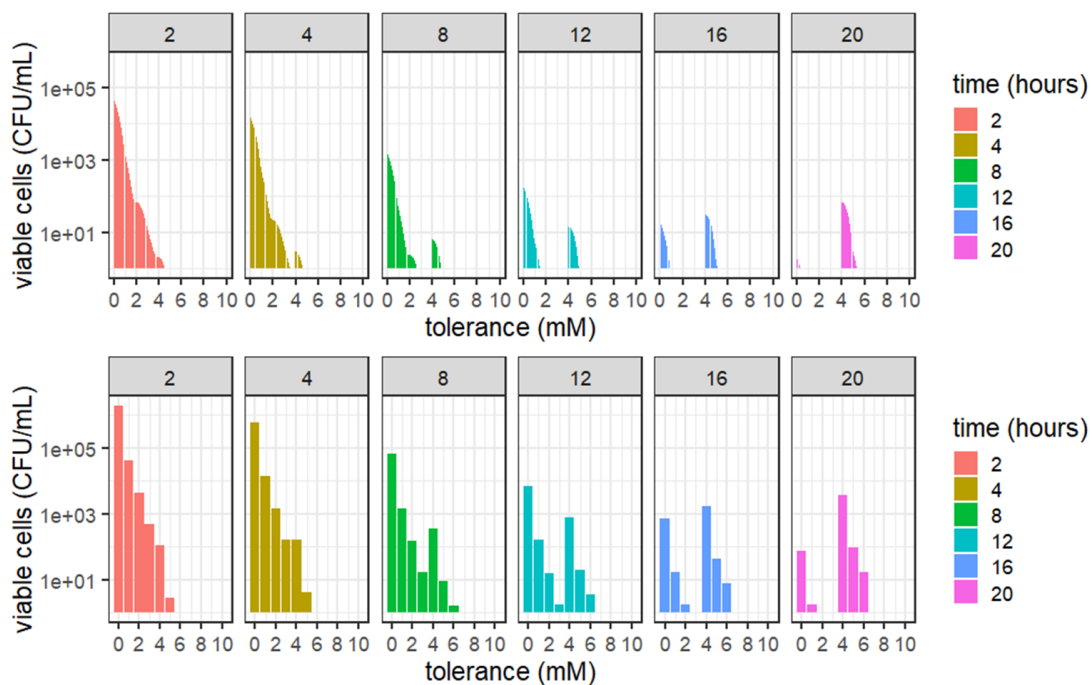


Figure 2.16 - Model result of change of tolerance distribution in growth with 4 mM formaldehyde, results from the absolute death version. Top: continuous results from numerical solving of PDE. Bottom: aggregated results of the continuous solution.

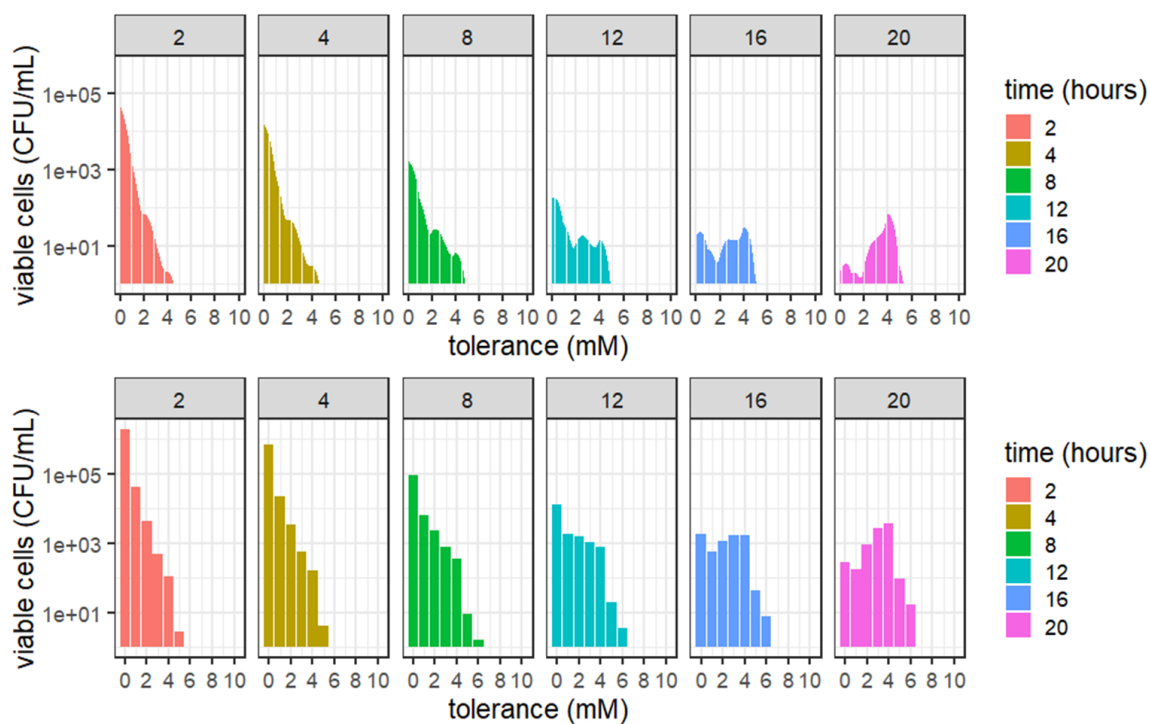


Figure 2.17 - Model result of change of tolerance distribution in growth with 4 mM formaldehyde, results from relative death version. Top: continuous results from numerical solving of PDE. Bottom: aggregated results of the continuous solution.

2.4.7 Likelihood ratio test and model selection

To determine whether there is an evidence for movement between phenotypic states, I asked whether growth and death were sufficient to explain the dynamics of the tolerance distribution, and if not, whether there was a support for either random, bidirectional movement (represented by a diffusion operator), or a directional movement (represented as advection). Different types of phenotypic movements are shown in Table 2.2.

Table 2.2 - Different phenotypic movements and corresponding mathematical implementation.

	Advection	Diffusion
Directional movement	+	-
Random movement	-	+
Both directional and random movement	+	+

Likelihood ratio test was performed to evaluate models. For likelihood ratio tests, models have to be nested; so models with one phenotypic parameter (either advection or diffusion) were compared with the null model (no phenotypic movement) and in each scenario (combination of initial conditions and death versions) any model that had significant p-value over the null model was compared with a model that has both diffusion and advection. The likelihood ratio is given by:

$$LR = -2(LL_0 - LL_1)$$

where LR is the likelihood ratio, LL_0 is the log likelihood from the reduced model and LL_1 is the log likelihood from a model with additional parameters (full model). Chi-squared tests were performed to evaluate the significance of LR with degrees of freedom given by the difference between the number of parameters in the full model and the reduced model.

2.4.8 AIC calculation

Likelihood ratio tests were used to determine the best combination of phenotypic movement parameters for each death function and initial condition combination. AIC values were used to compare the models with different death functions. Specifically, I compared the models with the best combination of phenotypic movements obtained from the likelihood ratio tests using AIC. AIC values were obtained from the log likelihood values (see 2.4.3):

$$AIC = 2(k - LL)$$

where k is the number of estimated parameters (Akaike, 1974). In each scenario, the AIC value for the null model (no phenotypic movement), a model with advection, a model with diffusion or a combination of both advection and diffusion was calculated. The best fit parameters are shown in Table 2.3.

Table 2.3 - Parameters and their values. Value shown by * shows the estimate for diffusion in the regrowth on succinate media.

Parameters	Description	Value	Units	Reference
K_m	Half concentration of methanol where k_{cat} of MDH is half maximum	0.02	mM	(Anthony and Zatman, 1964)
K_s	Half concentration of succinate where k_{cat} of succinate transporter is half maximum	0.003	mM	(McALLIS TER and Lepo, 1983)
K_f	Concentration of formaldehyde where k_{cat} of FAE is half maximum	0.2	mM	(Vorholt et al., 2000)
α	Formaldehyde dependent death rate	0.189	$mM^{-1}h^{-1}$	This study
r_m	Growth rate in methanol media	0.195	h^{-1}	This study
r_s	Growth rate in succinate media	0.267	h^{-1}	This study
V_{maxm}	Combined parameter of MDH concentration and its specific activity	2.59×10^{-8}	$mMmlh^{-1}cell^{-1}$	This study
V_{maxs}	Combined parameter of succinate transporter concentration and its specific activity	9.09×10^{-9}	$mMmlh^{-1}cell^{-1}$	This study
V_{maxf}	Combined parameter of FAE concentration and its specific activity	1.32×10^{-8}	$mMmlh^{-1}cell^{-1}$	This study
ν	Advection coefficient	0.269	mMh^{-1}	This study (estimated)
D	Diffusion coefficient	0.0278- 0.0233*	mM^2h^{-1}	This study (estimated)

2.5 Results

2.5.1 Bimodality of tolerance distribution during transition from net death to net growth suggests death is an absolute cutoff with tolerance level

The model was assessed to understand the role of death (Figure 2.10) in predicting changes in the tolerance distribution during growth on methanol with formaldehyde (Figure 2.7). Results from the PDE model for both absolute and relative death versions with estimated parameters from death curves and no phenotypic movement (advection or diffusion) were simulated and compared with the data. The absolute death version qualitatively better captured the disruptive distribution seen in the data; cell number in tolerance levels below 4 mM showed a rapid decline and 4 mM subpopulation increased in number that resulted in a bimodal distribution. In contrast, the relative death function maintained a single-peaked distribution that only gently declined below 4 mM. Because a cell's death rate is proportional to the difference between its tolerance level and the external concentration of formaldehyde, cells with tolerance just below 4 mM grew only slightly slower than those with tolerance above 4mM (Figure 2.18). Although the absolute death version captured the qualitative features of the data, without advection or diffusion, the bins at moderate tolerance levels (bins 1, 2, and 3 mM) were lower than seen in the data. This is particularly evident by looking to the 16 hours time-point (Figure 2.19).

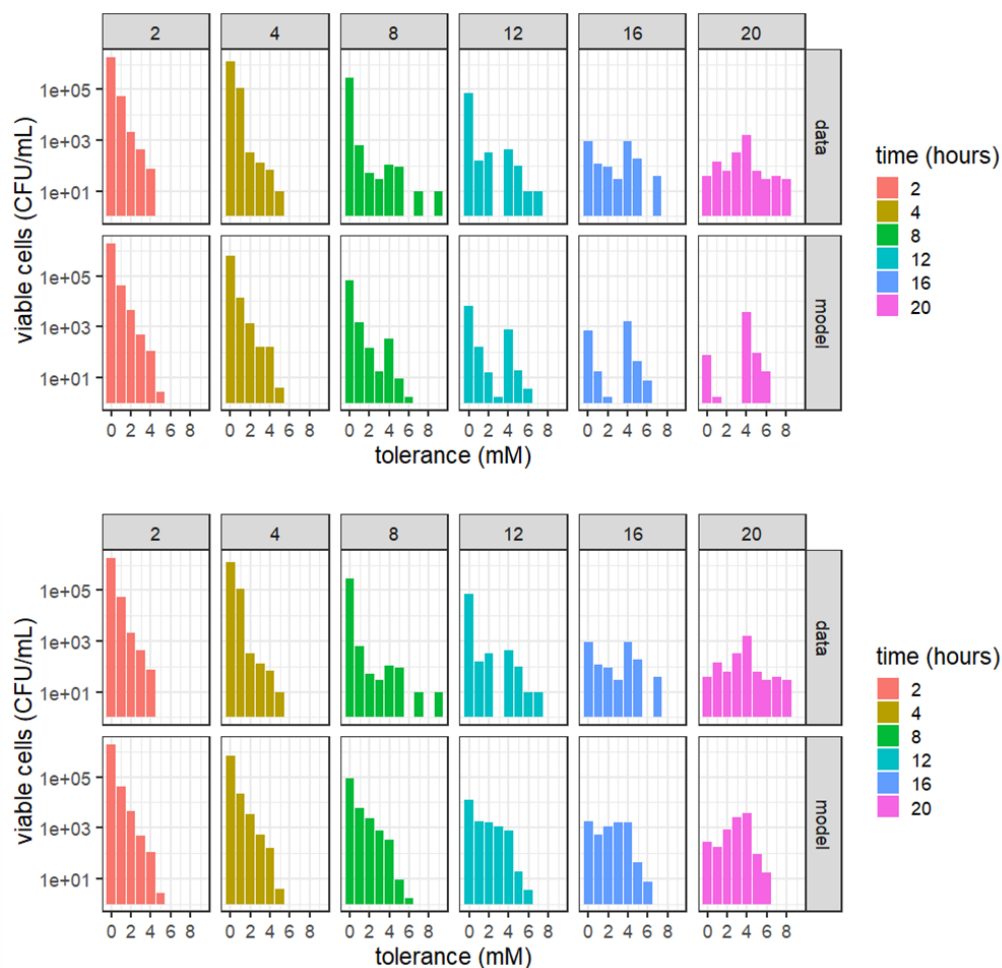


Figure 2.18 - Model results are aggregated into discrete bins to be compared with the data. Top: absolute death version. Bottom: relative death version. Absolute death version captures bi-modal peaks of the distribution (i.e. time-points 12 hours and 16 hours). Relative death version captures the spread between the bins.

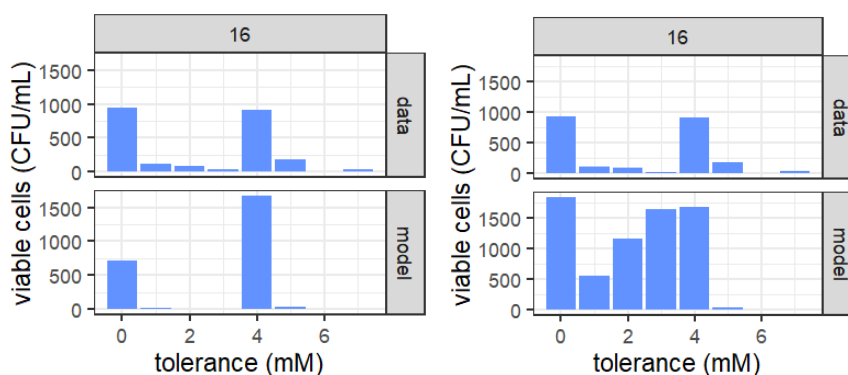


Figure 2.19 - Comparison between the result of the model and the data at time 16 hours in linear scale. Left: result of the absolute death version, the model is able to capture two peaks but not the spread between those. Right: result of the relative death version, the model is not able to capture two peaks, but in contrast to the absolute version, it is able to produce the spread between the bins.

2.5.2 Diffusion is sufficient to explain the tolerance shift in growth with formaldehyde

Data in growth on methanol with formaldehyde (Figure 2.7) suggest that in addition to growth and selection by death, there is a mechanism to spread the cells between different tolerance levels. To investigate the effect of adding phenotypic movements to different death functions, diffusion and advection parameters were estimated from the growth data in both death versions and two versions of initial conditions (without extension and with extension beyond the limit of detection). In all cases, the estimate for advection was very small. Estimation of parameters and model statistics are shown in Table 2.5. The absolute death model with added diffusion had the highest R^2 values. In case of the initial condition without extension, adding advection had a significant advantage (p-value: 0.0167); but in case of the initial condition with extension, adding advection did not make the model better (p-value: 0.194). The model with added diffusion and extended initial condition had the lowest AIC (223.82). The selected model for the growth on methanol with 4 mM formaldehyde is shown below (Equation 9):

$$\frac{\partial N(x, t)}{\partial t} = r_m N(x, t) - H(x, F)N(x, t) + D \frac{\partial^2 N(x, t)}{\partial x^2} \quad (9)$$

As it was shown in the results with no phenotypic movement, the relative death function lacked the bimodal distribution in tolerance, and the inclusion of diffusion did not change this result. In contrast, adding diffusion to the model with the absolute death function maintained the bimodal distribution and captured much of the spread in the middle tolerance bins (1, 2, 3 mM) seen in the data (Figure 2.20).

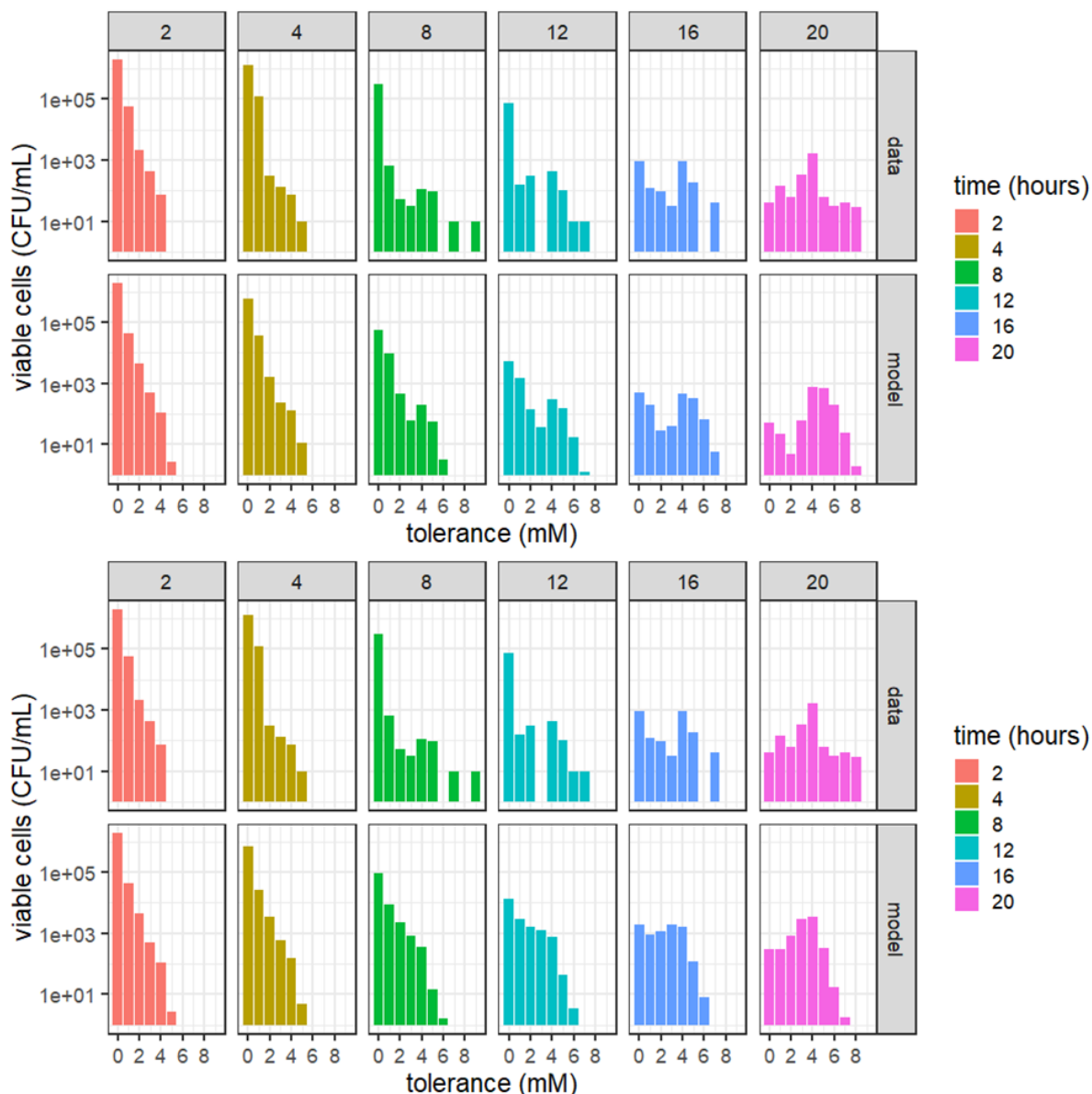


Figure 2.20 - Results of different death functions with diffusion included in the model. Top: the tolerance distribution of viable cells from the absolute death version of the model with diffusion. The distribution shows bimodality at 16 hours. Bottom: the tolerance distribution of viable cells from the relative death version of the model with diffusion. The distribution lacks bimodality at 16 hours and has too few highly tolerant cells at 16 and 20 hours.

2.5.3 Advection is necessary to explain the shifting back of tolerance in regrowth on succinate medium, but not methanol medium

Data on regrowth of selected tolerant subpopulation in succinate media (Figure 2.8) suggests there should be a mechanism for cells with high tolerance levels to lose their tolerance. In this section, I investigate the ability of the model without any phenotypic movement to explain the regrowth data on

methanol and succinate. Simulating regrowth in methanol (Figure 2.21, top, methanol/model) showed that growth is sufficient to match the data (Figure 2.21, top, methanol/data). Using the model to simulate the regrowth on succinate failed to capture the cells losing tolerance (Figure 2.21, top).

The diffusion and advection parameters were estimated separately using data from each condition (i.e., regrowth on methanol or succinate). Estimation of parameters and model statistics are shown in Table 2.6. In the methanol environment, the null model had the lowest AIC (98.38) and phenotypic movement was not significant. In the succinate environment, the model with both advection and diffusion had the lowest AIC (115.68). Equations 10 and 11 show the regrowth in methanol and succinate media respectively, since there is no formaldehyde and consequently no death in the regrowth experiment, the death terms are not included in the equations. The tolerance distribution for the best-fit models are shown in Figure 2.21.

$$\frac{\partial N(x, t)}{\partial t} = r_m N(x, t) \quad (10)$$

$$\frac{\partial N(x, t)}{\partial t} = r_s N(x, t) + v \frac{\partial N(x, t)}{\partial x} + D \frac{\partial^2 N(x, t)}{\partial x^2} \quad (11)$$

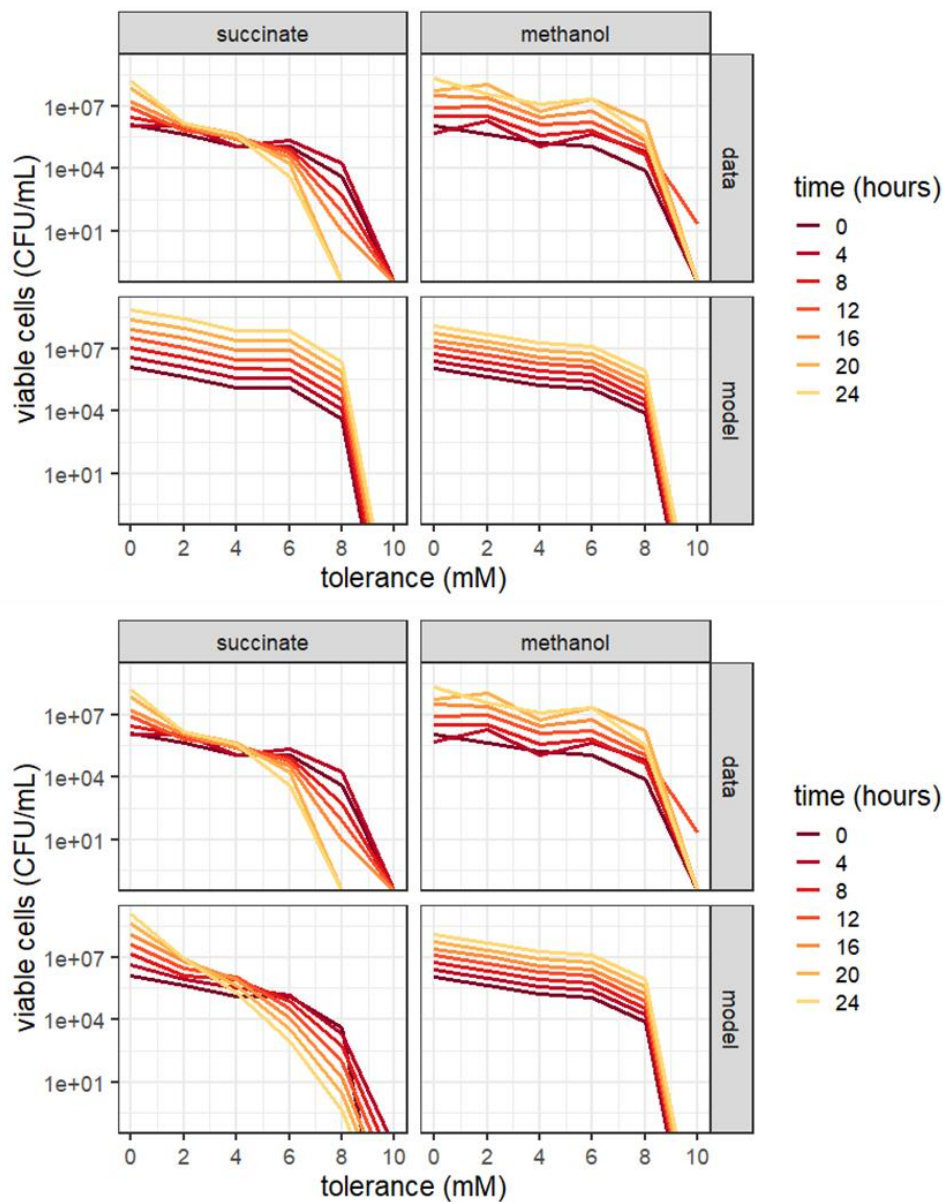


Figure 2.21 - Advection is necessary to capture the regrowth on succinate. Top: result of the model with no phenotypic movement; model lacks the losing tolerance in succinate media. Bottom: model results with added advection in the succinate media.

2.6 Discussion

The mathematical model developed in this project elucidates the effect of different death functions on shaping the tolerance distribution in growth on methanol treated with formaldehyde, the presence of a random bidirectional phenotypic movement (diffusion) in spreading cells to different tolerance levels, and the necessity for a mechanism to lose the tolerance in regrowth on the succinate media (advection). One of the key points from our model is that death rate shows a sharp threshold between tolerant and

sensitive cells. Here, I have considered two extreme versions of death functions, but many other mathematical forms could be proposed to relate a toxin's concentration to death rate. For example, a saturation version of death is common in modeling antibiotic death curves, also known as Zhi models (Zhi et al., 1986). As we have seen already, relative death rate looks similar to the effect of diffusion in spreading the tolerance, but it fails to generate a sharp bimodal curve.

The finding that the data are best approximated with an absolute death function, suggesting that there should be a sharp threshold between survival and death (Figure 2.3). This border between death and viability is discrete, although the threshold of formaldehyde that a given cell can withstand is continuous. This threshold may be due to the nature of formaldehyde toxicity. Formaldehyde is a potent damaging agent which, in contrast to antibiotics, does not have specific targets, and damages all macromolecules, including proteins. The other difference is that formaldehyde serves as a carbon source for cells. If a cell's tolerance level is high enough to overcome formaldehyde toxicity, they can grow. But if formaldehyde's consumption machinery is not able to overcome toxicity, formaldehyde damages macromolecules including enzymes that are necessary for its consumption. This situation generates a positive feedback loop, where cells that begin to fail to deal with formaldehyde toxicity will only become less able to manage this potent stressor (similar to work in antibiotic resistance, Deris et al., 2013).

In my model, I simply used x as the tolerance level as it manifests as a cellular phenotype, but it remains unclear what intracellular differences lead to x . Tolerance could be a simple function of the concentration of a single type of macromolecule in the cell or could depend upon many components in a more complex manner. Formaldehyde consumption involves more than one component (Marx et al., 2003), so tolerance level could be the joint effect of two or more macromolecules that we observe as one dimension of cellular phenotype in the data. One of the fundamental aspects of the model is the assumption of a continuous distribution of phenotypes. In this model, transitions between phenotypic states occur only locally; no jumps in phenotypes are permitted. This assumption is not necessarily true. If we assume a phenotypic state is the result of a regulatory event, the stochastic nature of gene regulation could lead to jumps in phenotype. Other model formulations such as integral projection models could account for more general transitions including jumps (Merow et al., 2014).

For changing the phenotypic state, I used simple and convenient (Perthame, 2015) mathematical forms: advection and diffusion. These processes have different interpretations, depending upon what the variable x represents (e.g., tolerance). If we consider the tolerance state to be the result of a macromolecule or a combination of macromolecules, diffusion could be seen as unequal inheritance of macromolecules by cells in a population. Advection could be the result of degradation or generation

of a macromolecule, or down-regulation or up-regulation of genes encoding macromolecules production.

In different environmental conditions, we saw different processes dominate (Table 2.4). Without formaldehyde, like regrowth on succinate, advection dominates diffusion. It is important to note that there should be always a baseline level of diffusion, as advection only would collapse the distribution in the lower end of tolerance space. In the medium formaldehyde case, like regrowth on methanol, cells maintain their tolerance level even after recovering from formaldehyde stress. This means that, although methanol alone is not shifting the distribution to higher levels, it is able to keep the distribution at high levels of tolerance. This is an example of hysteresis, where behavior of a population depends on its previous condition (Deris et al., 2013; Igoshin et al., 2008; Savageau, 1999). In the high formaldehyde scenario, diffusion dominates advection and cells bi-directionally move in tolerance space.

Table 2.4 - Different environmental conditions and corresponding population's response.

No formaldehyde	Medium formaldehyde	High formaldehyde
$v \gg D$	Maintaining tolerance level	$D \gg v$

2.7 Supplementary materials

2.7.1 Estimating parameters

2.7.1.1 *Estimating growth rate on succinate media*

The growth rate on succinate has been calculated from growth curves using simple linear regression. Figure 2.22 shows the experimental growth curves on 3.75 mM of succinate with the fitted growth rate. Figure 2.23 shows the residuals of the fit. Estimated growth rate r_s from time 7.5 to 16 is 0.267 ± 0.005 .

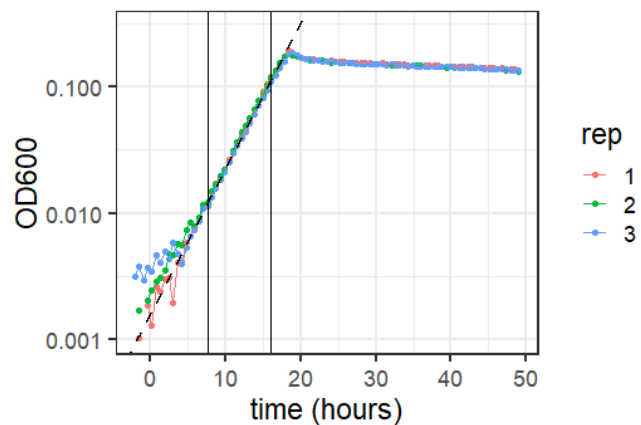


Figure 2.22 - Fitting growth rate on 3.75 mM succinate with three replicates. The distance between the two vertical lines was used to fit the growth rate; the dashed line shows the fitted line from the estimated growth rate.

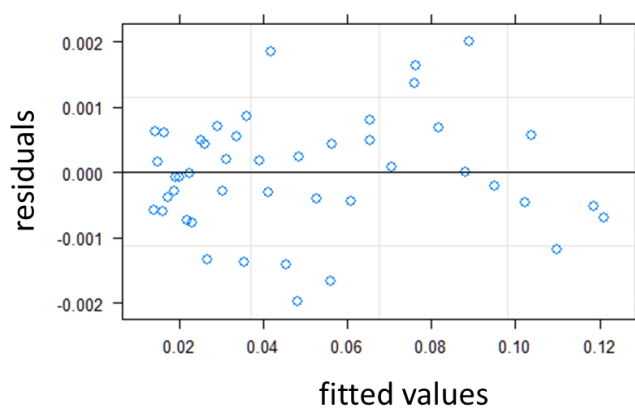


Figure 2.23 - Residuals of succinate growth fit.

2.7.1.2 Estimating growth rate in methanol media

As above for succinate media, the growth rate in methanol media can be calculated from growth curves. Figure 2.24 shows the experimental growth curves on 15 mM methanol and the fitted growth rate. Figure 2.25 shows the residuals of fitting. Estimated growth rate r_m is 0.195 ± 0.001 .

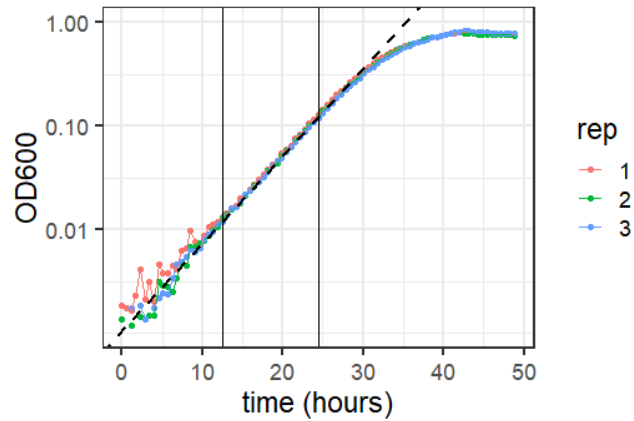


Figure 2.24 - Fitting the growth rate on 15mM methanol. The distance between the two vertical lines was chosen to fit the exponential growth rate. The dashed line shows the fit.

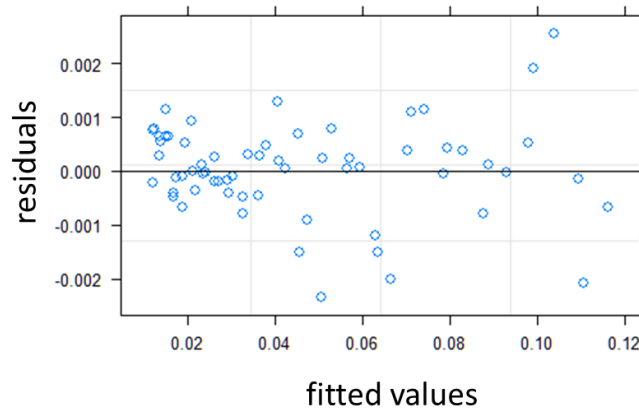


Figure 2.25 - Residuals of the methanol fit.

2.7.1.3 Calculation of V_{maxs}

In order to analytically solve for the succinate utilization rate as a function of the growth rate, I simplified the two equations of $\frac{dS(t)}{dt}$ and $\frac{dN(t)}{dt}$ to find V_{maxs} . For the exponential part of the growth, the succinate concentration is very high, thus we can approximate the $\frac{dS(t)}{dt}$ and $\frac{dN(t)}{dt}$ equations:

$$\frac{S(t)}{S(t) + K_s} \approx 1$$

So I can rewrite the $\frac{dS(t)}{dt}$ and $\frac{dN(t)}{dt}$ as follows:

$$\frac{dS(t)}{dt} = -V_{maxs} \frac{S(t)}{S(t) + K_s} N(t) \approx -V_{maxs} N(t)$$

$$\frac{dN(t)}{dt} = r_s \frac{S(t)}{S(t) + K_s} N(t) \approx r_s N(t)$$

The solution of $N(t)$ could be written as:

$$N(t) = N_0 e^{r_s t}$$

Plugging the solution above for $N(t)$ into the $\frac{dS(t)}{dt}$ equation and integrating yields:

$$\int dS(t) = \int -V_{maxs} N_0 e^{r_s t} dt$$

$$S(t) = S_0 - \frac{V_{maxs} N_0 e^{r_s t}}{r_s}$$

$$V_{maxs} = \frac{e^{-r_s t} r_s (S_0 - S(t))}{N_0}$$

In the equation above $S(t) = K_s = 0.003$, $S_0 = 3.75$, $r_s = 0.267$. To calculate N_0 , the initial OD₆₀₀ measurement was converted to viability by utilizing a linear relationship between OD₆₀₀ and viability. The relationship between OD₆₀₀ and viability is shown in Figure 2.26. From linear regression between the viability and OD₆₀₀ we have: $CFU = (5.17 \pm 0.10)10^8 \times OD_{600}$.

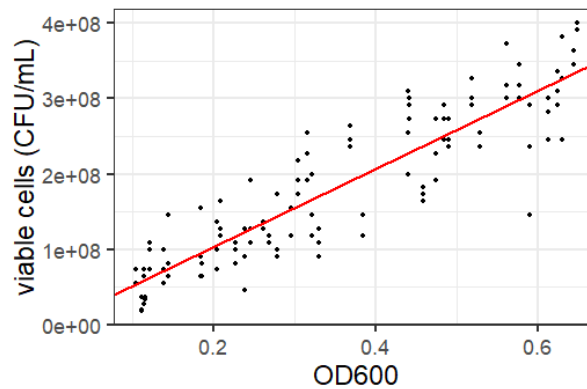


Figure 2.26 - OD₆₀₀ and number of viable cells show a linear relationship.

For calculating V_{maxs} we need to know the time that the culture reaches stationary phase (τ_s). This time interval was selected as when the OD₆₀₀ emerges above the noisy background levels to the point when the culture saturates (Figure 2.27). The time from the beginning of the culture to the stationary phase is τ_s . The initial noisy part of the growth was not considered as part of τ_s . The first vertical line is at 5.316 hours and the second line is at 18.371 hours. The difference between these two lines is $\tau_s = 18.371 - 5.316 = 13.055 \pm 0.567$.

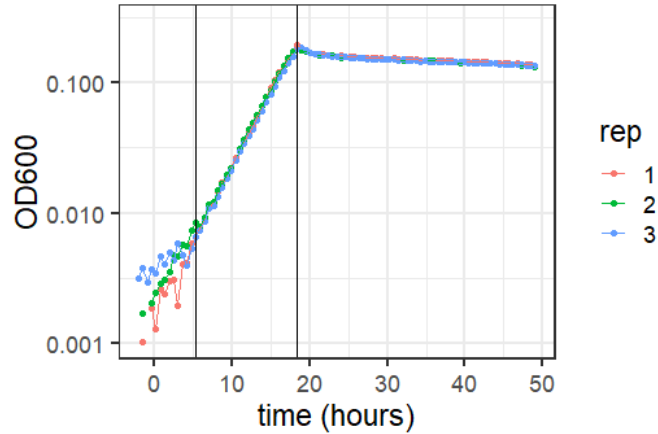


Figure 2.27 - Growth on 3.75 mM succinate. The time from the end of the noisy initial period of growth to the stationary phase is τ_s . The distance between the two vertical lines is τ_s

The initial OD_{600} is the mean of replicates 1 and 3 at the beginning of τ_s which is 6.527×10^{-3} . The initial population in terms of $CFU = 6.527 \times 10^{-3} \times 5.17 \times 10^8 \approx 3.37 \times 10^6$. In our equation, $t = \tau_s$. Solving the equation for V_{maxs} yields:

$$V_{maxs} = \frac{e^{-r_s t} r_s (S_0 - S(t))}{N_0}$$

$$V_{maxs} = \frac{e^{-0.267 \times 13.055} \times 0.267 (3.75 - 0.003)}{3.37 \times 10^6} \approx 9.09 \times 10^{-9}$$

Simulation of growth on 3.75 mM succinate, using the calculated value for V_{maxs} , is shown in Figure 2.28.

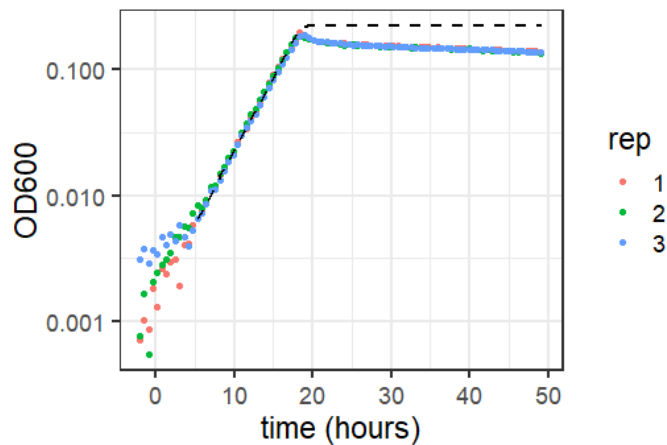


Figure 2.28 - Simulated growth on 3.75mM succinate. The dashed line shows the fit from the model

2.7.1.4 Calculation of V_{maxm}

Similar to the succinate equation, we can simplify and write the general solution of $M(t)$ as follows:

$$\frac{M(t)}{M(t) + K_m} \approx 1$$

$$\frac{dM(t)}{dt} = -V_{maxm} \frac{M(t)}{M(t) + K_m} N(t) \approx -V_{maxm} N(t)$$

$$\frac{dN(t)}{dt} = r_m \frac{M(t)}{M(t) + K_m} N(t) \approx r_m N(t)$$

which has solution

$$N(t) = N_0 e^{r_m t}$$

Plugging the solution for $N(t)$ into the $\frac{dM(t)}{dt}$ equation and integrating yields:

$$\int dM(t) = \int -V_{maxm} N_0 e^{r_m t} dt$$

$$M(t) = M_0 - \frac{V_{maxm} N_0 e^{r_m t}}{r_m}$$

$$V_{maxm} = \frac{e^{-r_m t} r_m (M_0 - M(t))}{N_0}$$

To calculate τ_m or time for consumption of methanol from the beginning of the experiment, this quantity equals to the t in the equation above. The issue with methanol growth data, however, is that methanol is volatile, and thus both depletes some wells, and allows others to continue to grow due to cross-well gas transfer. This prevents direct use of the time of consumption.

I also calculated τ_m using an alternative approach. If OD_{600} at the beginning of the succinate and methanol growth is the same, then the time for consumption of a substrate has an inverse relationship with the growth rate. Thus, I can write:

$$\tau_m = \tau_s \frac{r_s}{r_m}$$

This implies:

$$\tau_m = 13.057 \frac{0.267}{0.195} = 17.878$$

Thus, in the $M(t)$ equation above, $M(t) = K_m = 0.02$, and $M_0 = 15$. The mean of replicates 1 and 3 at time 9.646 hours is 6.681×10^{-3} , which is the nearest OD_{600} to the initial OD_{600} of the succinate culture. With these values, the N_0 in terms of the CFU is $6.681 \times 10^{-3} \times 5.17 \times 10^8 \approx 3.45 \times 10^6$, $t = \tau_m = 17.878$ and $r_m = 0.195$.

$$V_{maxm} = \frac{e^{-r_m t} r_m (M_0 - M(t))}{N_0}$$

$$V_{maxm} = \frac{e^{-0.195 \times 17.878} \times 0.195 (15 - 0.02)}{3.45 \times 10^6} = 2.59 \times 10^{-8}$$

Simulating growth on 15 mM succinate using the calculated value for V_{maxm} is shown in Figure 2.29.

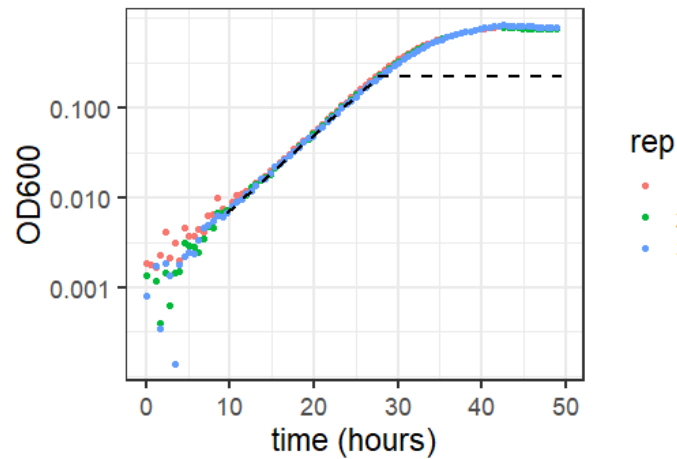


Figure 2.29 - Simulated growth on 15mM methanol. The dashed line shows the fit. In the data, cells grow for a longer period of time than in the model, as there is some residual methanol transferred between wells. The simulation result is consistent with a previous study (Delaney et al., 2013a).

2.7.1.5 Calculation of V_{maxf}

Since in all of the experiments involving formaldehyde, growth happens using methanol as a substrate,

I use the $\frac{dN(t)}{dt}$ equation for the methanol case as it was shown earlier:

$$\frac{dN(t)}{dt} = r_m \frac{M(t)}{M(t) + K_m} N(t)$$

Since $M(t)$ and $F(t)$ are not changing during most of the growth, I re-write the $\frac{dF(t)}{dt}$ equation as:

$$\frac{dF(t)}{dt} \approx -V_{maxf} N(t)$$

Plugging the $N(t)$ solution in the equation above yields:

$$\int dF(t) = \int -V_{maxf}N_0e^{r_mt} dt$$

$$F(t) = F_0 - \frac{V_{maxf}N_0e^{r_mt}}{r_m}$$

$$V_{maxf} = \frac{e^{-r_mt}r_m(F_0 - F(t))}{N_0}$$

In this equation, $F(t) = K_f = 0.2$, $F_0 = 1$, I need N_0 and the first positive OD₆₀₀ number is 0.002 at time 3.5 hours in the OD₆₀₀ plot (Figure 2.30).

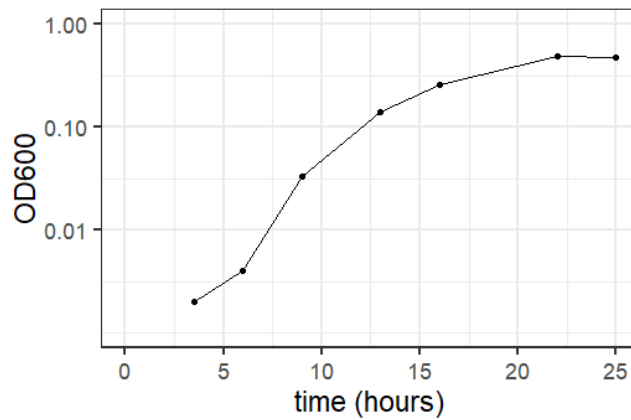


Figure 2.30 - Growth on 15 mM methanol treated with 1 mM formaldehyde. The first positive OD₆₀₀ number is at time 3.5 hours.

According to the formaldehyde assay (via Nash), it takes 16 hours for a culture to consume formaldehyde. The parameters were calculated as $N_0 = 0.002 \times 5.17 \times 10^8 = 1.034 \times 10^6$, $t = 16 - 3.5 = 12.5$, $r_m = 0.195$, and $V_{maxf} = \frac{e^{-0.195 \times 12.5} 0.195 (1 - 0.2)}{1.034 \times 10^6} = 1.32 \times 10^{-8}$.

Using these parameter values, simulation of consumption of formaldehyde and experimental data are shown in Figure 2.31.

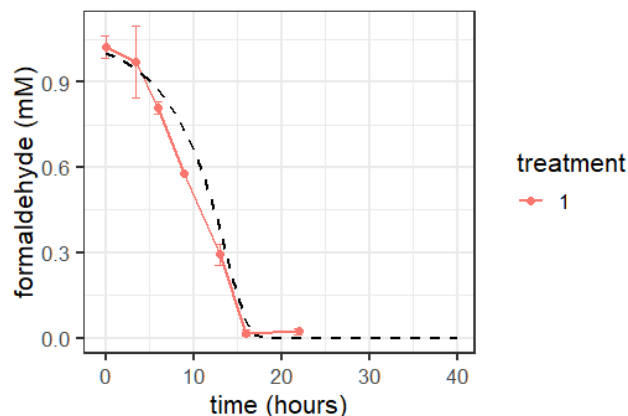


Figure 2.31 - Formaldehyde measurement data and fit.

2.7.1.6 Alternative way of calculating V_{maxf}

According to the experimental data (Figure 2.32) it takes 84 hours for formaldehyde to decrease from 4 mM (F_0) to 0.12 mM.

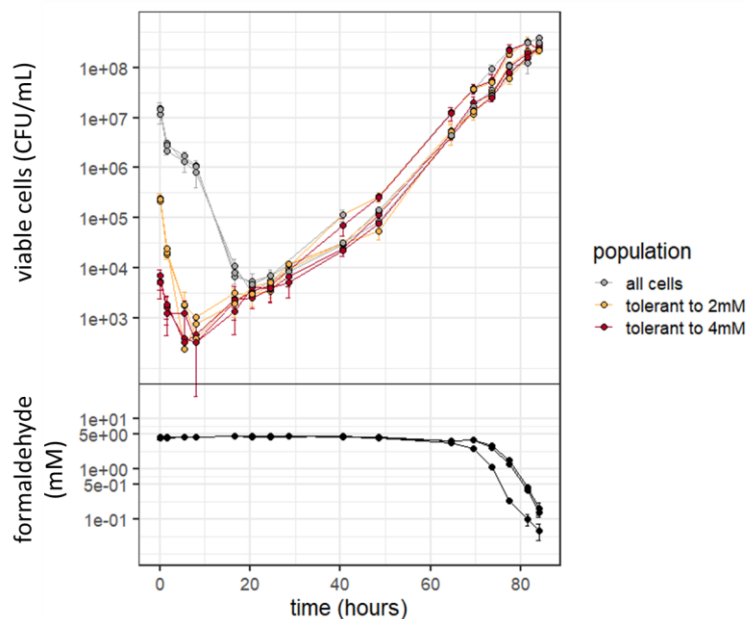


Figure 2.32 - Viability and Nash assay data for growth on 4mM formaldehyde. Top: viability over time for different tolerant populations. Bottom: Nash assay data for consumption of formaldehyde

After 24 hours, all of the cells are tolerant to 4 mM formaldehyde. The mean of the three replicates for all cells at time 24.5 is 6626.263. The time for consumption of formaldehyde, starting from this point is: $84 - 24.5 = 59.5$. Thus, V_{maxf} can be written as:

$$V_{maxf} = \frac{e^{-r_m t} r_m (F_0 - F(t))}{N_0} = \frac{e^{-0.195 \times 59.5} \times 0.195 (4 - 0.12)}{6626.263} \approx 1.04 \times 10^{-9}$$

The modeled formaldehyde concentration using the calculated V_{maxf} (1.04×10^{-9}) is shown in Figure 2.33.

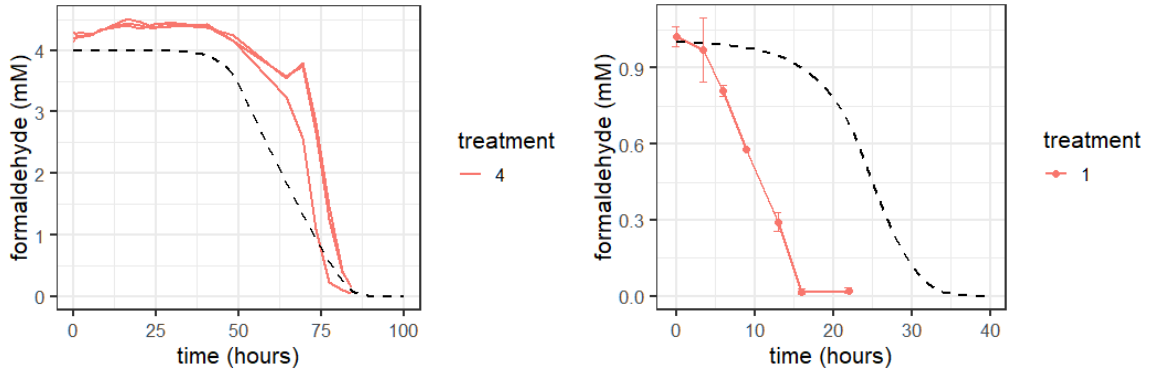


Figure 2.33 - Formaldehyde measurement data (red) and simulated result using estimated V_{maxf} (dashed line). Left: 4 mM formaldehyde measurement data. Right: 1 mM formaldehyde measurement data.

Comparison of the result with V_{maxf} calculated from the 1 mM formaldehyde measurement (1.32×10^{-8}) is shown in Figure 2.34.

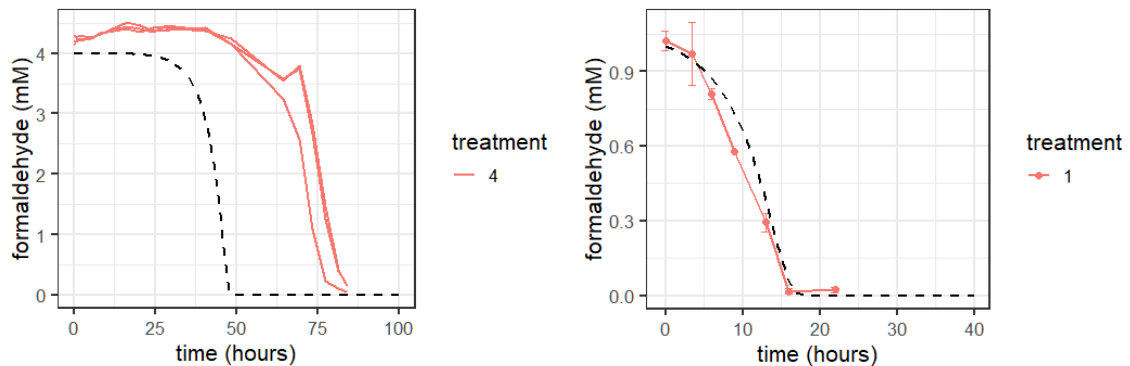


Figure 2.34 - Formaldehyde measurement (red) and simulated result using estimated V_{maxf} (dashed line). Left: 4 mM formaldehyde measurement data. Right: 1 mM formaldehyde measurement data.

2.7.1.7 Estimation of diffusion and advection forward on growth data

The diffusion and advection parameters for the growth experiment, separately or in combination, were estimated, and log likelihood, R^2 and p-values were calculated for different death functions and initial conditions. Values are shown in Table 2.5.

Table 2.5 - Estimation of phenotypic movement parameters (diffusion and advection) for the growth in 4 mM formaldehyde scenario. Estimates of the parameters and their standard errors, log likelihood (LL), AIC, R^2 values, and p-values calculated from the likelihood ratio test are shown in the table below. The filled cell with blue color shows the best model.

Initial condition	Parameter choice	Absolute death rate	Relative death rate
IC without extension	Null model	LL= -140.68, AIC=281.36, $R^2=0.820$	LL= -140.97, AIC=281.93, $R^2=0.819$
	Estimating v only	$v=3.31E-3 \pm 3.24E-3$, D=0, LL= -137.15, AIC=276.29, $R^2=0.840$, p-value vs null = $7.84E-3$	$v=4.22E-2 \pm 1.33E-2$, D=0, LL= -139.28, AIC=280.56, $R^2=0.829$, p-value vs null = $6.63E-2$
	Estimating D only	$v=0$, $D=3.15E-2 \pm 5.38E-3$, LL= -112.80, AIC=227.61, $R^2=0.929$, p-value vs null = $8.22E-14$	$v=0$, $D=5.71E-3 \pm 2.11E-3$, LL= -130.41, AIC=262.82, $R^2=0.872$, p-value vs null = $4.33E-6$
	Estimating v and D	$v=8.35E-2 \pm 3.57E-2$, $D=6.36E-2 \pm 1.80E-2$, LL= -109.94, AIC=223.88, $R^2=0.936$, p-value vs D = $1.67E-2$	$v=2.06E-1 \pm 5.97E-2$, $D=8.45E-2 \pm 3.31E-2$, LL= -122.60, AIC=249.20, $R^2=0.902$, p-value vs D = $7.76E-5$
IC with extension	Null model	LL= -128.74, AIC=257.49, $R^2=0.879$	LL= -129.32, AIC=258.63, $R^2=0.877$
	Estimating v only	$v=2.91E-3 \pm 2.09E-3$, D=0, LL= -123.40, AIC=248.81, $R^2=0.899$, p-value vs null = $1.08E-3$	$v=2.44E-2 \pm 1.89E-2$, D=0, LL= -128.58, AIC=259.16, $R^2=0.880$, p-value vs null = $2.25E-1$
	Estimating D only	$v=0$, $D=2.78E-2 \pm 6.40E-3$, LL= -110.91, AIC=223.82, $R^2=0.933$, p-value vs null = $2.34E-9$	$v=0$, $D=4.25E-3 \pm 3.60E-3$, LL= -128.07, AIC=258.13, $R^2=0.882$, p-value vs null = $1.14E-1$
	Estimating v and D	$v=5.20E-2 \pm 3.83E-2$, $D=5.00E-2 \pm 1.85E-2$, LL= -110.07, AIC=224.14, $R^2=0.935$, p-value vs D = $1.94E-1$	$v=1.90E-1 \pm 6.30E-2$, $D=7.77E-2 \pm 3.42E-2$, LL= -123.09, AIC=250.19, $R^2=0.900$, p-value vs D = $1.61E-3$

2.7.1.8 Estimation of diffusion and advection backward on regrowth data

For the re-growth experiment, diffusion and advection parameters were estimated separately or in combination, and log likelihood, R^2 and p-values were calculated for either the methanol or the succinate case. Values are shown in Table 2.6.

Table 2.6 - Estimation of phenotypic movement parameters (diffusion and advection) for the re-growth scenario. Estimates of the parameters and their standard errors, log likelihood (LL), AIC, R^2 values, and p-values calculated from the likelihood ratio test are shown in the table below. The filled cells with blue color show the best models.

Parameter choice	Methanol	Succinate
Null model	LL= -49.19, AIC=98.38, $R^2=0.996$	LL= -110.06, AIC=220.13, $R^2=0.81$
Estimating v only	v= 1.31E-3±5.60E-4, D=0, LL= -49.11, AIC=100.23, $R^2=0.996$, p-value vs null = 1.00	v= 1.63E-1±3.80E-3, D=0, LL= -68.44, AIC=138.88, $R^2=0.981$, p-value vs null = 0.00
Estimating D only	v= 0, D=5.18E-6±3.00E-4, LL= -49.19, AIC=100.38, $R^2=0.996$, p-value vs null = 9.69E-1	v= 0, D=4.34E-19±8.18E-4, LL= -110.06, AIC=222.13, $R^2=0.812$, p-value vs null = 1.00
Estimating v and D	v= 5.17E-4±6.26E-4, D=1.97E-10±2.97E-4, LL= - 49.16, AIC=100.32, $R^2=0.996$, p-value vs v = 1.00	v= 2.69E-1±2.61E-2, D=2.33E-2±7.96E-3, LL= - 56.84, AIC=115.68, $R^2=0.990$, p-value vs v = 1.47E-6

3 Analyzing Gene Expression to Understand the Response of *M. extorquens* to the Toxicity of Formaldehyde

3.1 Introduction

Stressors in biological systems can have multiple layers of action. They can have single or numerous molecular targets, each of which may affect different cellular processes, and ultimately the consequences can radiate throughout the whole cell. Antibiotics, for example, inhibit particular proteins, which themselves are involved in the synthesis of protein, DNA or cell walls (Kohanski et al., 2010). Ultimately, this leads to a slowing or cessation of growth. On the other hand, stressors such as heat, osmotic pressure, radioactive radiation, pH change, metals, or aldehydes do not have specific targets; all of these examples cause proteins to misfold. Although there are still layers to the response – the proteins that are misfolded, and the processes affected by them – both levels are broad, with many proteins and processes being affected simultaneously.

In *Methylobacterium extorquens*, an internal stressor – formaldehyde – is generated during the growth on single-carbon compounds such as methanol. Formaldehyde is generated and consumed as a central intermediate at a rate of ~ 2 mM/s (Vorholt et al., 2000). *M. extorquens* is able to tolerate formaldehyde in low concentrations, but at high concentrations formaldehyde is toxic to the cell (Marx et al., 2003). Furthermore, growth of wild-type *M. extorquens* PA1 on formaldehyde is only possible at concentrations of ~ 1 mM.

Experimental evolution of *M. extorquens* PA1 on ever-increasing concentrations of formaldehyde uncovered a novel protein, EfgA, which appears to be critical to the stress-response system for formaldehyde (Nayak et al., in prep). All three replicate populations ultimately grew on 20 mM formaldehyde, and from genome sequencing and subsequent targeted sequencing, it became clear that the first mutation in all cases occurred in the DUF336 domain of a gene with an unknown function. DUF336 domains mainly occur in genes of unknown function, but there is one characterized homolog (HbpS) that is a sensor of oxidative damage (Bogel et al., 2009; de Orué Lucana et al., 2009; Ortiz de Orué Lucana et al., 2016). This gene was named *efgA* for enhanced formaldehyde growth. It was immediately noted that close homologs to *efgA* are exclusively found in methylotrophic bacteria, cementing the idea that this gene is ecologically relevant to their specific metabolism. Genetic analyses revealed that the evolved *efgA* alleles were all loss-of-function mutations, and thus even deleting *efgA* permitted growth on formaldehyde. On the other hand, evidence for a physiological benefit of EfgA was found when methanol grown cells were shocked with high concentrations of formaldehyde (30

mM); here, the $\Delta efgA$ strain had decreased survival when compared to the WT, demonstrating it was beneficial during acute formaldehyde stress.

Subsequent analysis of EfgA revealed that it acts by directly binding formaldehyde, which causes it to interact with peptide deformylase and inhibit translation. A second round of experimental evolution to formaldehyde growth, with a larger number of populations, provided clues that EfgA may interact with the ribosome. While the majority of these populations happened upon additional *efgA* alleles, genome sequencing revealed that three isolates had beneficial mutations in *def*, which encodes peptide deformylase (PDF). PDF is an essential gene and is required for the processing of the majority of peptides produced by the ribosome (Adams, 1968). Pull-down assays with FLAG-PDF with His₆-EfgA confirmed that EfgA binds formaldehyde and directly interacts with PDF in a formaldehyde-dependent manner.

The interaction between formaldehyde and EfgA results in translation inhibition, thus its action has similarity to translation-inhibiting antibiotics, such as kanamycin (Figure 3.1, top). How then can EfgA be beneficial to cells? To investigate the translation-inhibiting role of EfgA, we used RNA sequencing (RNA-seq) analysis to compare the changes in global expression patterns between WT and $\Delta efgA$ mutant when treated with formaldehyde or, in parallel, kanamycin. Because the combination of EfgA and formaldehyde, or kanamycin alone, have a common function (i.e., translation inhibition), many downstream consequences of translation inhibition may be in common. These two processes also have notable differences: they each involve different specific players and in addition, formaldehyde also acts as a general, multi-target stressor and a potential carbon source.

In an experiment led by Dr. Jannell Bazurto, three replicate populations of WT *M. extorquens* and the $\Delta efgA$ mutant were treated independently with formaldehyde, kanamycin and no-stressor, as a control. Over an 18 hour timecourse, various parameters (viability, cell density and external formaldehyde concentration) were tracked and samples were collected for further analyses (Figure 3.1, bottom). Our collaborators in the laboratory of Dr. Jeffrey Barrick at the University of Texas, Austin performed rRNA depletion and RNA-seq on selected samples from the earlier time points, which included pre-treatment (45 minutes before adding the treatment) and 5, 20, 40, 180 and 360 minutes post-treatment. For mapping the RNA reads to specific genes, the Bowtie2 2.2.6 alignment tool was used with the *Methylobacterium extorquens* PA1 genome (GenBank: CP000908.1) as the reference sequence. Counting of reads was carried out using HTSeq 0.6.1p1 (Anders et al., 2015).

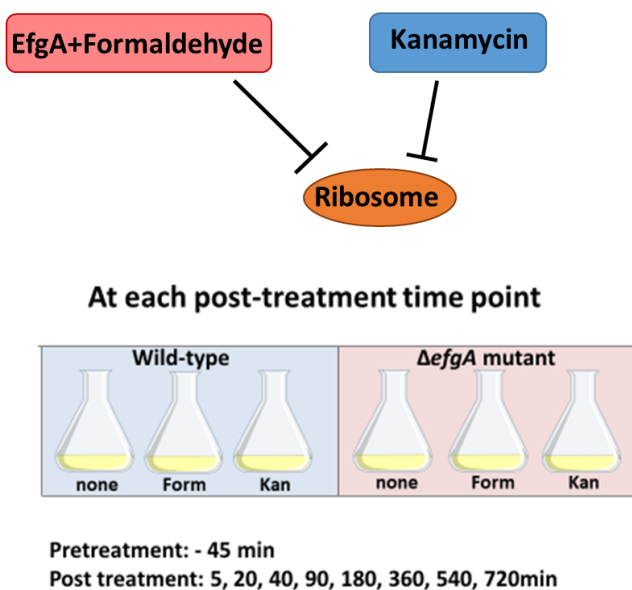


Figure 3.1 - Common role of kanamycin and formaldehyde in translation inhibition and design of the stress exposure experiment. Top: like kanamycin, the combination of the EfgA protein with formaldehyde inhibits translation by interacting with the ribosome. Bottom: isogenic populations of WT and Δ *efgA* were grown in succinate (15 mM) minimal medium in biological triplicates. During early exponential phase of growth, cultures were treated with formaldehyde (Form, 5 mM), kanamycin (Kan, 50 μ g/mL) or left untreated (None). The untreated cultures served as the no-stressor control for comparison. Cells were monitored for up to 18 hours; RNA sequencing was performed on samples taken from the indicated time points.

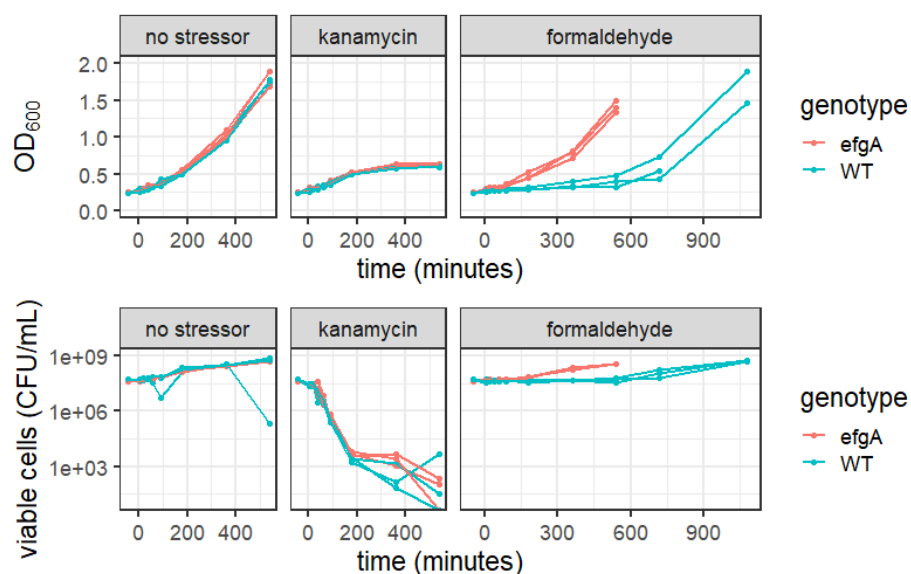


Figure 3.2 - Comparison of the growth response of WT (green) and Δ *efgA* mutant (red) in three different conditions: no-stressor, kanamycin and formaldehyde. When present, stressors (kanamycin and formaldehyde) were introduced into growth media at time = 0 minute. The top row shows optical density measured at 600 nm (OD₆₀₀) and bottom row shows cell viability measured in colony forming units per milliliter (CFU/mL).

Critical for interpreting the changes in global gene expression is the fact that formaldehyde (with EfgA) and kanamycin exhibited very different effects upon growth and viability over the timecourse of this experiment. In the no-stressor treatment, viability and OD measurements showed an equivalent increase for both genotypes, suggesting that, in the absence of formaldehyde, the absence of EfgA does not alter cell growth. For both genotypes, the addition of kanamycin did not impact the OD until 180 minutes (Table 3.4). By contrast kanamycin treatment quickly resulted in the loss of viability, which dropped for both genotypes by an order of magnitude in the first 40 minutes, and continued to fall rapidly until 180 minutes. Despite the introduction of formaldehyde to the media, the $\Delta efgA$ strain grew nearly as well as the no-stressor control, as measured by OD and viability. On the other hand, by 40 minutes, the OD of WT showed a significant difference in OD from the control by 40 minutes (Table 3.4), and had not increased by 180 minutes (Figure 3.2). In summary, the response to kanamycin showed a similar behavior in OD of both genotypes, only slowing long after viability had fallen, whereas in the formaldehyde treatment, WT showed an immediate delay in OD without a loss in viability increment versus $\Delta efgA$ OD behaving nearly like the no-stressor condition.

In this chapter, I investigate changes in gene expression using RNA-seq data to understand how the cells responded to different forms of translation inhibition. First, I describe the overall patterns in global expression for the WT or $\Delta efgA$ strains that are differentially stressed (no-stressor, kanamycin, or formaldehyde). Next, I investigate the degree of overlap or uniqueness between conditions or genotypes. Finally, I narrow my analysis to assess the roles of specific genes in the formaldehyde response. Specifically, these genes were either candidate genes previously identified as having a role in the formaldehyde response or they were identified in overrepresented categories of functional genes that showed significant changes in expression in this experiment.

3.2 Methods

3.2.1 Normalization of the data

All data manipulation and statistical analysis was done in R. The matrix of raw count data was converted to normalized counts using DESeq2 package under a negative binomial model (Love et al., 2014). The package does the normalization using the method of “median of ratios” (Anders and Huber, 2010).

3.2.2 Principal Component Analysis and heatmaps

For Principal Component Analysis, the `plotPCA()` function in DESeq2 was used. For heatmap plots, normalized counts were calculated across all replicates and then the means of replicates were

calculated. Heatmaps were generated using the `heatmap.2()` function in the `gplots` package. For the heatmaps, rows representing genes were clustered using the `hclust()` default function. All of the treatments were compared to WT pre-treatment (the control pre-treatment with no stressor at the 45 minutes before adding any treatment). To calculate the normalized counts for the heatmap plots, I used this formula to avoid problems encountered with zero counts:

$$counts^* = \log_2 \left(\frac{counts + 1}{counts\ in\ WT\ pre_treatment + 1} \right)$$

3.2.3 Venn diagrams

For the Venn diagrams, all of the treatments for WT and $\Delta efgA$ genotypes were compared with their own pre-treatment (WT pre-treatment and $\Delta efgA$ pre-treatment, respectively) to indicate whether a significant change had occurred. The Wald test was used to calculate the significance. To reduce the number of false positive genes across the ~5000 genes in the genome, I used a very conservative criterion: subsets of genes with False Discovery Rate (FDR) adjusted p-values less than 0.001 were selected, and based on sign of fold change, genes were divided into up or down-regulated categories.

3.2.4 Analysis of significance in candidate genes

For the modest number of candidate genes examined, the Wald test was used to calculate the significance. Genes with FDR adjusted p-values less than 0.05 were classified as significantly changed genes. Furthermore, for the formaldehyde metabolism genes, I used a one-tailed test because we had a specific direction for our *a priori* hypothesis. Genes with both positive fold changes and positive Wald-statistics were classified as up-regulated and those with negative fold changes and negative Wald-statistics were classified as down-regulated.

3.3 Results

3.3.1 Overall pattern of gene expression changes revealed via Principal Component Analysis (PCA)

As an initial step to assess the consistency between replicates and the overall trends in the data, I used principal components analysis (PCA). First, it was clear from the analysis that the replicates for each treatment were very well-clustered (Figure 3.3). Second, the pre-treatment timepoints, all of the no-stressor treatments, and the 40 minute timepoints for kanamycin treatment all clustered on top of each other. Even the later timepoints for the formaldehyde treatments ($\Delta efgA$ at 180, 360 minutes; WT at 720 minutes) fell into this cluster. Third, PC1 appears to have captured the strength of the shared

response seen for both kanamycin and formaldehyde as stressors. Fourth, PC2 appears to separate the formaldehyde treatment from the kanamycin one.

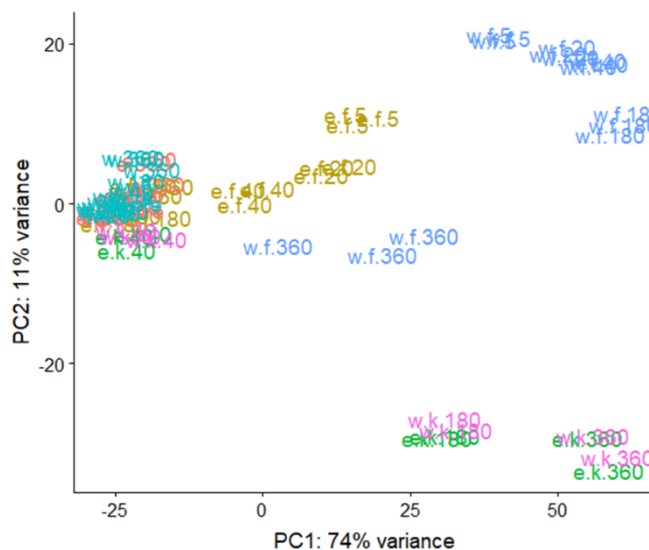


Figure 3.3 - PCA plot of all treatments in all timepoints in both WT and $\Delta efgA$. Control samples are in turquoise and orange for WT and $\Delta efgA$, formaldehyde treatments are in blue and gold, and kanamycin treatments are in pink and green. The labels indicate the genotype (w = WT; e = $\Delta efgA$), stressor (none for control; k = kanamycin; f = formaldehyde), and timepoint in minutes. All of the kanamycin late timepoints are close together (lower right). In WT treated with formaldehyde, cells show a high perturbation from early timepoints; at 360 minutes, the cluster is close to the pre-treatment. In $\Delta efgA$, like WT, we see a deviation from and coming back to the pre-treatment, but the perturbation is smaller compared to the WT.

3.3.2 Growth of WT and $\Delta efgA$ are similar to each other in the no-stressor condition

Beyond differences that may emerge due to the addition of stressors, formaldehyde and kanamycin, it was first critical to assess how much difference there was in gene expression profile between WT and $\Delta efgA$. In comparing the pre-treatment timepoint for each genotype, there was little indication of large-scale differences in fold-change of transcripts (Figure 3.4). There were only 8 (0.2%) genes that have significantly changed expression in $\Delta efgA$ pre-treatment compared to the WT pre-treatment. These data suggested that, in the absence of formaldehyde, the presence of EfgA has a minimal impact on global gene expression.

To determine if the no-stressor gene expression profiles of each strain were comparable for the duration of the experiment, I compared the RNA-seq of the two genotypes in no-stressor condition. Over the timecourse of the no-stressor treatment, only 0.4% of genes in WT and 0.2% of genes for $\Delta efgA$ showed a significant change at any time from 5 minutes to 180 minutes (Figure 3.5, left). Most of the

changes observed occurred at 360 minutes for both genotypes. There were 3.9% up-regulated and 1.3% down-regulated genes in WT at 360 minutes. In $\Delta efgA$, 5.3% and 3.8% of genes were up-regulated and down-regulated, respectively. From all the up-regulated genes at 360 minutes, 48.7% of up-regulated genes and 49.2% of down-regulated genes were in common between the two genotypes at 360 minutes (Figure 3.5, right). These data imply that growth with no stressor had a relatively constant expression pattern through time. Furthermore, this indicates that, during these conditions, there is relatively little impact of the absence of EfgA on global gene expression. These findings allow us to simplify the analyses with stressors described below by simply comparing expression profiles to the pre-treatment version of each genotype.

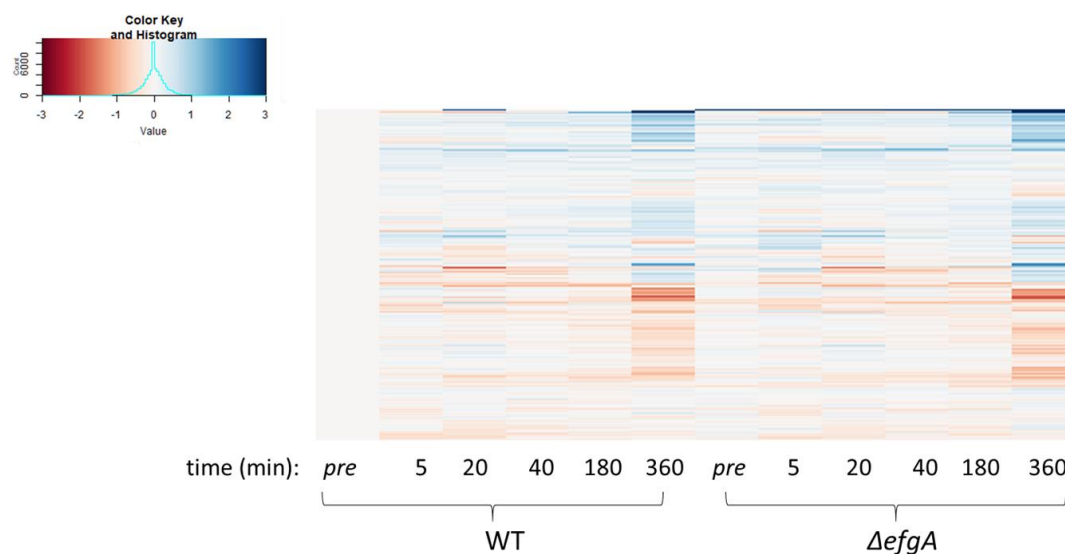


Figure 3.4 - Temporal heatmap plot of average fold change of gene expression data in the growth on succinate with no-stressor, for WT and $\Delta efgA$ strains. Rows represent genes and columns show timepoints for the two different genotypes. The data are \log_2 of normalized counts divided by normalized counts in WT pre-treatment (see methods). Blue shows up-regulated genes and red indicates down-regulated genes.

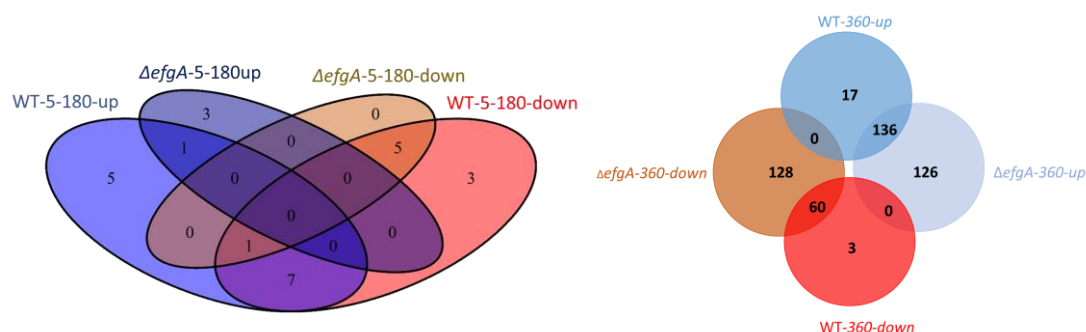


Figure 3.5 - Venn diagrams of differentially expressed genes in WT and $\Delta efgA$ with no-stressor over time. Left: “up” categories show all genes that have been up-regulated at any timepoint from 5-180 minutes with no-stressor and “down” categories show down-regulated genes from that time interval based on the sign of \log_2 fold change. Blue and purple show up-regulated genes in WT and $\Delta efgA$. Red and brown show down-regulated genes. Overall there are very few genes in both genotypes that show change in expression compared to the pre-treatments. Right: up and down-regulated genes in 360 minutes between WT and $\Delta efgA$. $\Delta efgA$ shows more up and down-regulated genes compared to the WT.

3.3.3 Treatment with kanamycin showed a delay in expression response compared to loss of viability

In order to assess the gene expression profile generated by the addition of a classical translational inhibiting antibiotic, kanamycin, both genotypes were analyzed at 40, 180, and 360 minutes after the addition of kanamycin. Although there had already been an order of magnitude drop in viability for each genotype by 40 minutes (Figure 3.2), the gene expression profiles each exhibited relatively modest changes at that time. It was only by 180 and 360 minutes that major changes in expression were observed (Figure 3.6). In WT, 30.1% and 40.6% of genes were differentially expressed at 180 minutes and 360 minutes, respectively (Figure 3.7). Similarly, in the $\Delta efgA$ strain 27.5% and 37.5% of genes showed significant changes in expression at 180 minutes and 360 minutes, respectively. For both strains the response at 180 minutes timepoint was largely a subset of the response that is amplified by 360 minutes. From all the up and down-regulated genes in both time points, only 10.7% of up-regulated genes and 7.2% of down-regulated genes were specific to 180 minutes in the WT strain. In $\Delta efgA$ 10.8% of up-regulated genes and 7.6% of down-regulated genes are only specific to 180 minutes.

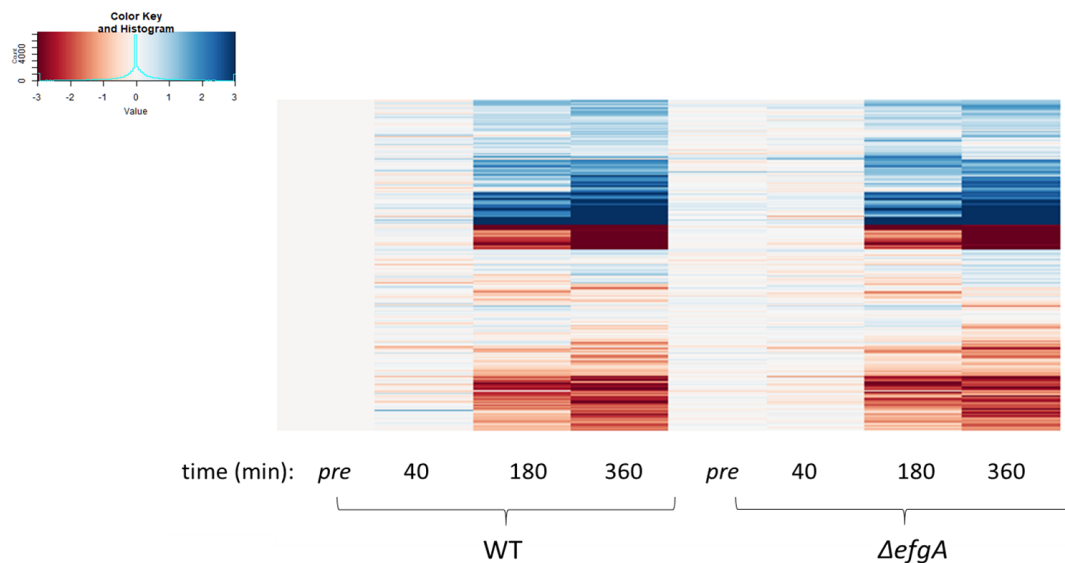


Figure 3.6 - Temporal heatmap of both genotypes in treatment with kanamycin. The heatmap shows intense responses to kanamycin happen at 180 minutes and continues to 360 minutes, while many genes are involved in this time interval when viability of cells are declining (Figure 3.2)

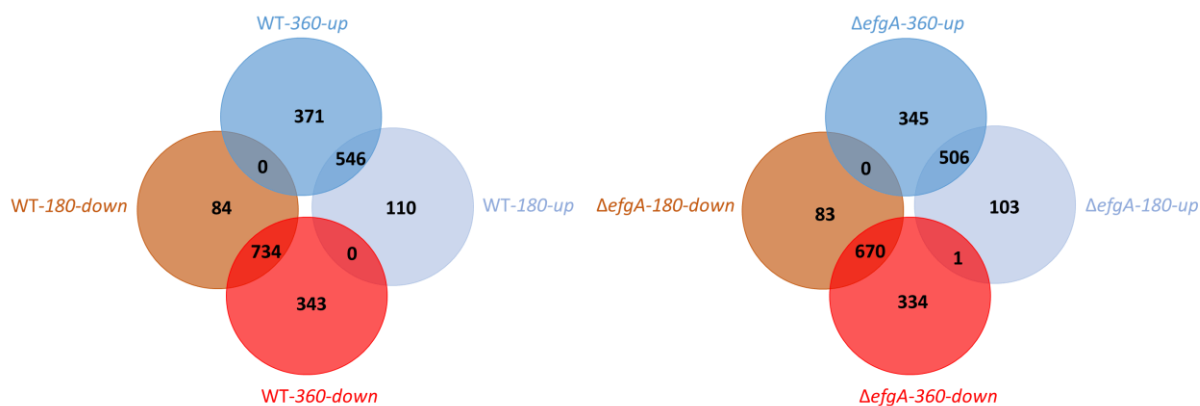


Figure 3.7 - Venn diagrams of differentially expressed genes at 180 and 360 minutes post kanamycin treatment. Left: WT and right: $\Delta efgA$ mutant. Genes that are up-regulated (up) and down-regulated (down) at 180 minutes are a subset of those at 360 minutes.

Since the response of cells in both genotypes at 360 minutes involved more genes, the overlap of differentially expressed genes at this timepoint was investigated (Figure 3.8). Of all the up-regulated genes, 80.8% of them were shared between the two genotypes and in the down-regulated set, 81.8% of all genes were shared between the two genotypes at 360 minutes. Additionally, not a single gene that went up in one genotype went down in the other (or vice versa). These data suggest that the presence of EfgA does not have a substantial impact on the response of cells to kanamycin-induced stress. As mentioned, the action of kanamycin is independent from formaldehyde and its interaction with EfgA; the presence of EfgA may not inhibit kanamycin from its translation inhibition function.

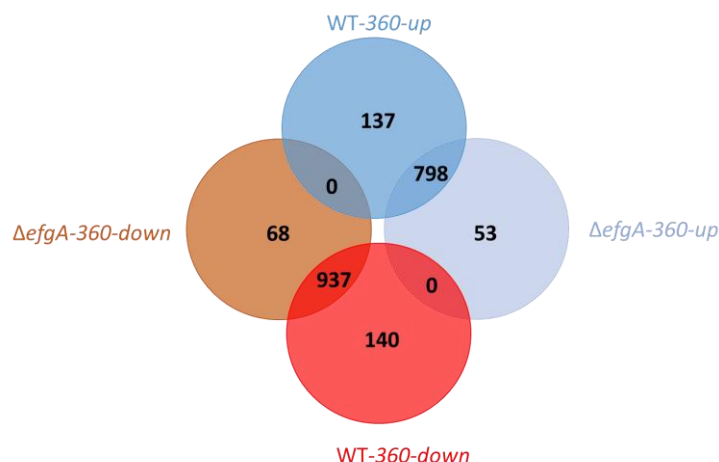


Figure 3.8 - Venn diagram of up and down-regulated genes in the kanamycin treatment at 360 minutes. Most of the genes are common between the two genotypes and there is no gene that is up-regulated in one genotype and down-regulated in the other genotype.

3.3.4 EfgA is the key component in response to formaldehyde

To investigate the role of EfgA in response to formaldehyde, temporal data from both genotypes treated with formaldehyde were analyzed. For WT, expression of 43.0% of genes were affected at 5 minutes, and changes continued to 360 minutes where 18.6% of genes showed significant changes. Overall, the response tended to increase up to 180 minutes, and then relaxed by 360 minutes (Figure 3.9), which was when external formaldehyde was below 1 mM (data not shown) and the OD had begun to increase. For $\Delta efgA$, 13.5% of genes changed expression at 5 minutes and 20.6% of genes showed significant change at 20 minutes. Even though more genes were involved at 20 minutes compared to the 5 minutes, the response became weaker; the average \log_2FC of up-regulated genes at 5 minutes was 2.37 and at 20 minutes was 1.90. For down-regulated genes the average at 5 minutes was -1.77 and at 20 minutes was -1.39. In both genotypes the genes showed a response at 5 minutes were a subset of the genes at 20 minutes (Figure 3.10).

Compared to kanamycin, which showed little response at 40 minutes and an increasing response by 360 minutes, with formaldehyde the pattern was quite different. The strongest response was seen immediately after the stress, and rapidly faded with later timepoints. It was also quite clear that the response in WT – with an active EfgA – was substantially greater than that in the $\Delta efgA$ mutant. The key feature in the response to formaldehyde was that, for WT, the response got stronger from 5 minutes to 180 minutes, and the response in each timepoint is a super-set of its previous time. The $\Delta efgA$ response, on the other hand, got weaker from 5 minutes to 180 minutes, but again each time point was largely a sub-set of its previous time point in terms of intensity (Figure 3.10). This pattern is in contrast

to examples of biological systems that exhibit waves of distinct gene expression with time, such as those seen for various classes of genes involved in either sporulation or flagellar synthesis (Keijser et al., 2007; Kim et al., 2016).

Even though the temporal dynamics and magnitude of response to formaldehyde differed between WT and $\Delta efgA$, we wished to determine how many of the genes involved were shared between the two genotypes. As an example, for the 5 minutes timepoint, 62.8% of the up-regulated genes and 75.9% of the down-regulated genes were only observed for the WT strain. Only 4.6% of the genes up-regulated in $\Delta efgA$ and 0.6% of the down-regulated ones were unique to $\Delta efgA$, the rest were shared with the WT strain (Figure 3.11). These data suggest that formaldehyde by itself is not the main component for this response, and that EfgA has a primary role. It also appears that the effect of formaldehyde in the absence of EfgA largely occurs in WT, too.

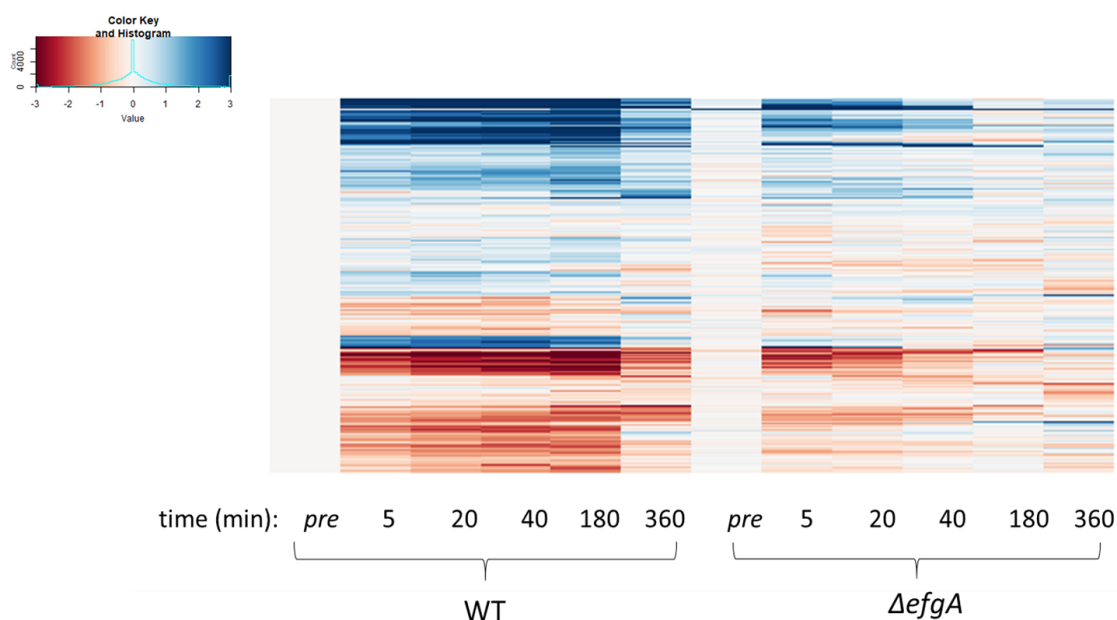


Figure 3.9 - Temporal heatmap plot of the two genotypes treated with formaldehyde. The data show an increasing response from 5 minutes to 180 minutes in WT, $\Delta efgA$ shows a response at 5 minutes in similar set of genes as WT but the response attenuates short after this timepoint.

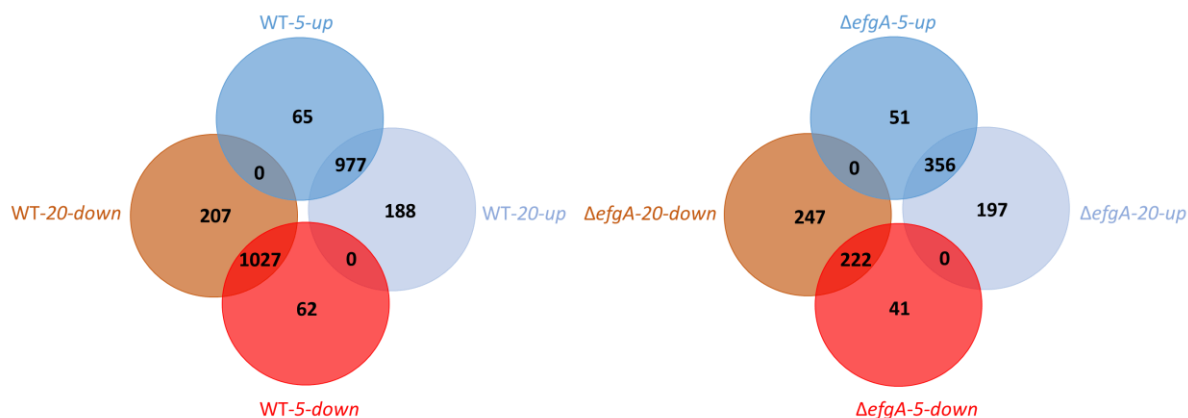


Figure 3.10 - Comparisons between 5 minutes and 20 minutes in treatment with formaldehyde. Left: WT and right: $\Delta efgA$. In both genotypes the response at 5 minutes is a subset of the response at 20 minutes timepoint.

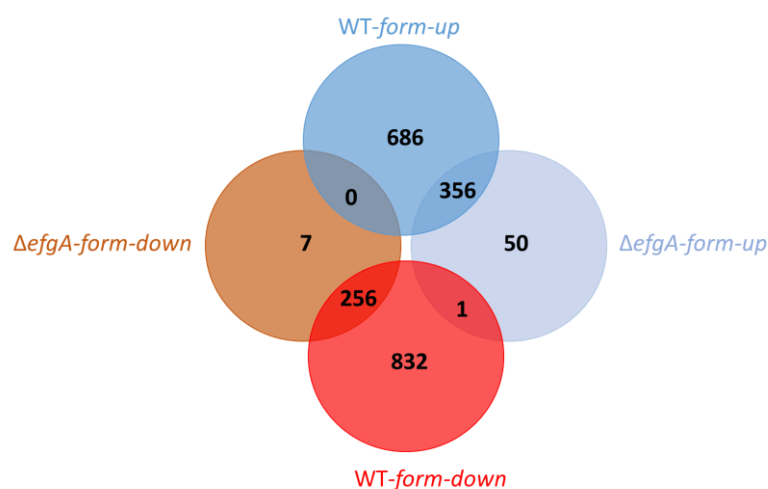


Figure 3.11 - Venn diagram of up/down regulated genes at 5 minutes with formaldehyde. Venn diagram shows majority of genes are specific to WT.

3.3.5 Response to kanamycin and formaldehyde involve shared pathways

Although EfgA+F and kanamycin both lead to translational arrest, the distinct differences in viability versus OD increase suggested that there may be very different responses involved. To assess the similarities and differences between these responses, we chose to compare the timepoints for WT with the strongest overall responses to each stress: formaldehyde at 5 minutes, and kanamycin at 360 minutes. Remarkably, 46.3% of the total up-regulated genes and 51.9% of down-regulated ones were in common between these two treatments (Figure 3.12). Since the outcomes of viability are so different, it is likely that many of these common genes represent the consequences of inhibited

translation. If this is the case, the key differences that are causal to how EfgA interacts with formaldehyde has an immediate, but non-lethal pause of translation are likely in the unique genes, or those that experienced opposite directions of change between the two stressors.

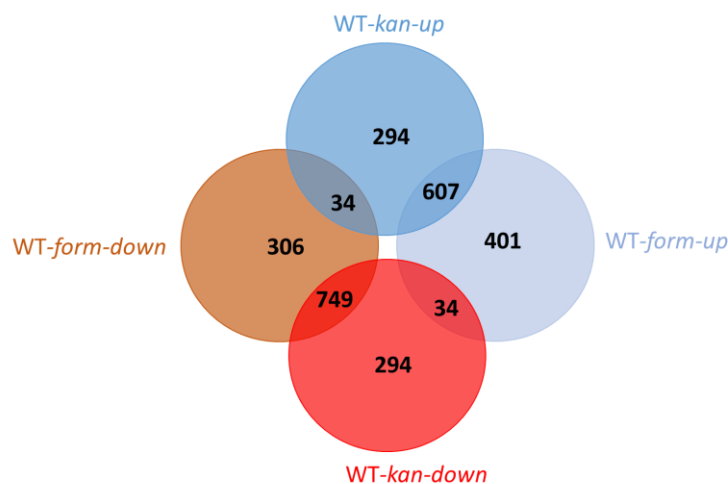


Figure 3.12 - Comparison between up/down-regulated genes in formaldehyde at 5 minutes and kanamycin at 360 minutes in WT. Even though there are many treatment-specific genes, majority of genes are shared between the two treatments, implying that responses to formaldehyde and kanamycin share many commonalities.

3.3.6 Formaldehyde oxidation genes showed down-regulation in treatment with formaldehyde

Formaldehyde is a central intermediate during metabolism of methanol (or other reduced C_1 compounds) as a carbon source, so changes in expression of genes involved in formaldehyde production and oxidation were expected when cells were exposed to formaldehyde stress. Specifically, I hypothesized that formaldehyde production (i.e., methanol oxidation) would be turned down, whereas formaldehyde oxidation (and perhaps utilization) would be turned up. To assess these changes, I compared genes involved in formaldehyde metabolism in both WT and $\Delta efgA$ treated with formaldehyde for 5 minutes compared to WT pre-treatment or $\Delta efgA$ pre-treatment, respectively. Genes involved in methanol oxidation, formaldehyde oxidation and formate oxidation are listed in Table 3.1. The *mx* operon, which encodes methanol dehydrogenase and is responsible for production of formaldehyde from methanol, was down-regulated. In contrast, most of the downstream enzymes showed up-regulation. The pattern was consistent between the two genotypes for most of the genes, as expected, since the EfgA mechanism of action does not involve formaldehyde metabolism. The genes *fhc* and *fdh3* showed down-regulation in WT but not $\Delta efgA$. The expression pattern showed that

cells up-regulated genes for consumption of formaldehyde, which should result in decreasing the cells' internal formaldehyde concentration (Figure 3.13).

Table 3.1 - Expression changes for genes involved in formaldehyde metabolism in both WT and $\Delta efgA$ at 5 minutes after exposure to formaldehyde. For each genotype, the first column shows the \log_2 FoldChange (\log_2 FC) compared to its pre-treatment condition, the second column shows the Wald test statistics (stat) and the third column shows the FDR-adjusted p-values (padj). Genes with significant change in expression are highlighted with color. Significantly up-regulated genes are shown in blue and down-regulated genes in orange. These data showed that most of the genes encoding enzymes involved for consumption of formaldehyde were up regulated, whereas the methanol dehydrogenase responsible for production of formaldehyde is down-regulated in both genotypes.

		WT			$\Delta efgA$		
Function	Gene	\log_2 FC	stat	padj	\log_2 FC	stat	padj
Methanol oxidation	<i>mxkB</i>	-2.5	-14.27	1.13E-44	-2.49	-9.14	2.73E-18
	<i>mxhH</i>	-3.27	-15.77	2.66E-54	-2.55	-7.67	5.23E-13
	<i>mxhE</i>	-2.1	-8.74	1.90E-17	-2.02	-5.59	3.03E-07
	<i>mxhD</i>	-1.46	-7.03	1.09E-11	-1.78	-6.79	2.44E-10
	<i>mxhL</i>	-3.49	-16.92	2.22E-62	-2.38	-6.9	1.13E-10
	<i>mxhK</i>	-2.52	-8.03	6.87E-15	-1.38	-3.46	3.23E-03
	<i>mxhC</i>	-2.97	-12.13	1.47E-32	-1.79	-6	3.01E-08
	<i>mxhA</i>	-3.02	-14	5.14E-43	-1.63	-5.36	1.01E-06
	<i>mxhS</i>	-2.64	-14.74	1.46E-47	-1.65	-5.48	5.38E-07
	<i>mxhR</i>	-2.42	-12.99	3.41E-37	-1.43	-5.39	8.87E-07
	<i>mxhI</i>	-0.86	-3.67	5.87E-04	-1	-3.76	1.17E-03
	<i>mxhG</i>	-2.68	-6.24	1.91E-09	-2.18	-5.09	3.99E-06
	<i>mxhJ</i>	-1.83	-10.3	8.94E-24	-0.96	-4.85	1.29E-05
	<i>mxhF</i>	-0.54	-2.12	5.61E-02	-0.28	-1	5.36E-01
<i>mxhW</i>	0.04	0.16	9.01E-01	0.26	0.78	6.52E-01	
Formaldehyde oxidation	<i>fae</i>	-0.17	-0.68	5.79E-01	0.52	2.34	6.77E-02
	<i>mtdB</i>	1.09	6.71	9.57E-11	1.45	5.75	1.28E-07
	<i>mch</i>	-0.07	-0.28	8.25E-01	0.47	1.96	1.44E-01
	<i>fhcC</i>	-0.88	-3.56	8.64E-04	0.32	1.03	5.20E-01
	<i>fhcD</i>	-1.23	-6.55	2.72E-10	-0.08	-0.31	8.76E-01
	<i>fhcA</i>	-0.99	-5.34	3.40E-07	-0.03	-0.13	9.51E-01
	<i>fhcB</i>	0.2	1.29	2.64E-01	0.4	1.65	2.43E-01
	<i>fdh1B</i>	-0.2	-1.03	3.81E-01	-0.88	-3.1	9.72E-03

Formaldehyde dehydrogenase	<i>fdh1A</i>	-1.07	-6.7	1.06E-10	-0.83	-3.61	1.96E-03
	<i>fdh2C</i>	6.21	20.3	2.45E-89	5.72	15.52	5.29E-52
	<i>fdh2B</i>	4.82	18.98	2.34E-78	5.15	16.29	2.71E-57
	<i>fdh2A</i>	2.26	12.1	2.14E-32	3.63	15.09	3.74E-49
	<i>fdh3A</i>	-0.18	-1	3.93E-01	0.25	1.18	4.52E-01
	<i>fdh3B</i>	-0.65	-3.96	1.97E-04	-0.08	-0.33	8.67E-01
	<i>fdh3C</i>	-0.5	-3.4	1.53E-03	-0.08	-0.39	8.43E-01
	<i>fdh4B</i>	1.9	6.61	1.90E-10	3.34	11.46	1.75E-28
	<i>fdh4A</i>	2.71	14.57	1.62E-46	3.45	13.96	4.13E-42
Formate assimilation	<i>ftfL</i>	2.15	14.56	1.97E-46	2.91	11.37	4.61E-28
	<i>fch</i>	1.16	5.74	3.78E-08	2.92	11.46	1.73E-28
	<i>mtdA</i>	1.89	9.72	2.56E-21	3.32	13.4	6.87E-39

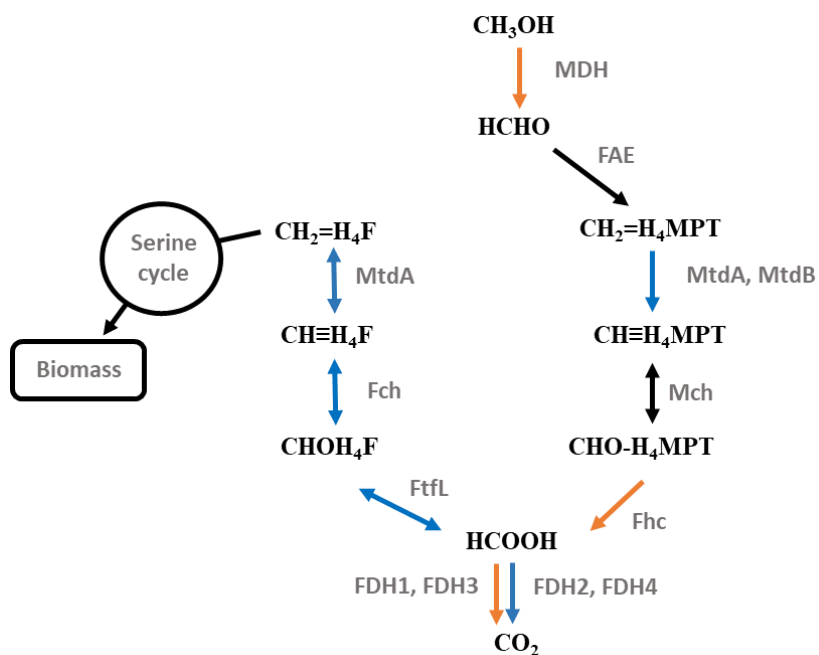


Figure 3.13 - Expression changes in genes encoding enzymes involved in formaldehyde metabolism. Up-regulated genes are shown in blue and down-regulated genes are shown in orange according to the Table 3.1. Most of the genes encoding enzymes involved in consumption of formaldehyde show up-regulation. The *mxs* gene cluster, encoding MDH responsible for production of formaldehyde showed down-regulation.

3.3.7 Few loci with beneficial mutations during formaldehyde evolution showed a significant change in expression upon formaldehyde exposure

I next examined whether loci where beneficial mutations occurred during the experimental evolution on increasing concentration of formaldehyde exhibited either increased or decreased expression upon formaldehyde exposure in either genotype. These genes were: *efgA* (as well as an *efgA* homolog), *efgB* (a predicted adenylyl cyclase), *def* (encodes peptide deformylase), another *def*-like homolog, and a *marR*-like gene (homolog of multi-antibiotic resistance regulator) were all investigated (Table 3.2). All of these genes were compared to the pre-treatment and Δ *efgA* pre-treatment respectively. In WT, there was a 3.7-fold increase in *efgA* (p-value = 1.04×10^{-7}). Additionally, there was a marginally significant 1.5-fold and 1.3-fold up-regulation in the *efgA* homolog (p-value= 5.01×10^{-2}) and the *def* homolog (p-value= 5.38×10^{-2}), respectively. In the Δ *efgA* strain, there was a 2.6-fold and 2.2-fold increase in *efgB* (p-value = 9.77×10^{-9}) and the *efgA* homolog (p-value= 1.03×10^{-2}). Additionally, there was a 1.5-fold increase in the *marR*-like gene (p-value= 7.65×10^{-2}).

Table 3.2 - Changes in expression upon formaldehyde exposure in genes that harbored beneficial mutations during formaldehyde evolution. For each genotype, the first column shows the \log_2 FoldChange (at 5 minutes) compared to its pre-treatment condition, the second column shows the Wald test statistics and the third column shows the FDR-adjusted p-values. Close homologs to *efgA* and *def* were included due to their uncertain biological role. Only *efgA* in WT and *efgB* and the *efgA* homolog in Δ *efgA* showed significant changes compared to their pre-treatments.

Gene	WT			Δ <i>efgA</i>		
	\log_2 FC	stat	padj	\log_2 FC	stat	padj
<i>efgA</i>	1.89	5.56	1.04E-07	-	-	-
<i>efgA</i> homolog	0.6	2.17	5.01E-02	1.11	3.08	1.03E-02
<i>efgB</i>	0.17	1.02	3.81E-01	1.5	6.2	9.77E-09
<i>def</i> (pdf)	0.21	0.86	4.70E-01	0.25	0.69	6.97E-01
<i>def</i> homolog	0.42	2.14	5.38E-02	-0.11	-0.36	8.56E-01
<i>marR</i>	0.3	1	3.92E-01	0.62	2.28	7.65E-02

3.3.8 Formaldehyde-induced genes involved in response to other common stresses

In addition to our *a priori* hypotheses regarding formaldehyde metabolism and beneficial loci in the experimental evolution, it is important to know what other functional genes have been affected the

most by treatment with formaldehyde. Looking at all of the annotated genes (excluding hypothetical proteins) in WT treated with formaldehyde at 5 minutes compared to WT pre-treatment, there were four classes of genes for which each constitutes at least 0.9% of all the annotated genes. Among the up-regulated genes, two groups that were seen repeatedly were chaperones and ABC transporters (Table 3.3). For down-regulated genes, there were again ABC transporters, as well as flagellar proteins and cytochromes. DNA damage proteins were both up and down-regulated. Given the overall pattern of partial overlap between formaldehyde and kanamycin responses, this list can be trimmed down to the exclusively formaldehyde-responsive genes by excluding those which changed during the kanamycin treatment. Finally, to assess the impact of EfgA activity, we can further remove those genes seen in the $\Delta efgA$ strain.

Looking at the subset of genes which are not in common with the kanamycin response at 360 minutes showed that part of the response in these frequent up/down-regulated groups were shared in response to kanamycin as well. In addition, excluding annotated genes that were in common with $\Delta efgA$ treated with formaldehyde showed part of these frequent up/down-regulated genes showed change in expression in $\Delta efgA$ as well, suggesting the response to formaldehyde is not dependent on only the EfgA protein and there should be other factors involved in sensing and responding to formaldehyde and its effects on physiology besides EfgA.

Table 3.3 - Total number of annotated genes in functional gene groupings that were differentially expressed in WT treated with formaldehyde (5 minutes). Column F shows total genes changed and directionality of changes is indicated by “Up” and “Down”. Column F(-kan) shows the subset of changed genes which is specific to formaldehyde translation inhibition (i.e., excludes genes common to kanamycin response at 360 minutes). Right side of the table shows number of genes that are only exclusive to WT and not in common with $\Delta efgA$. Chaperones were up-regulated. Flagellar proteins and cytochromes were mostly down-regulated. ABC transporters and DNA damage repair proteins were both up and down-regulated.

	WT				Exclusive to WT (not in $\Delta efgA$)			
	Up		Down		Up		Down	
	F	F(-kan)	F	F(-kan)	F	F(-kan)	F	F(-kan)
Total	687	315	888	232	432	218	666	183
Chaperones	13	4	1	0	2	0	0	0
DNA damage proteins	6	1	6	3	4	0	4	3
ABC transporter	24	15	22	5	10	5	18	5
Flagellar proteins	1	1	45	7	1	1	4	0
Cytochromes	5	3	17	3	4	2	11	3

3.3.9 The response of $\Delta efgA$ compared to WT involved general stress response proteins

Assessing the changes in beneficial loci (*efgA*, *efgB*, *def* and a *marR*-like gene) showed that *efgB* had higher expression in the $\Delta efgA$ mutant than WT in formaldehyde treatment; in order to find other genes that showed stronger response in formaldehyde relative to WT, I compared significant up/down-regulated genes in $\Delta efgA$ treated with formaldehyde at 5 minutes compared to WT treated with formaldehyde at 5 minutes. In order to limit this gene set to only those which responded more strongly in formaldehyde and remove the genotype effect, these set of genes were also compared with $\Delta efgA$ pre-treatment and only subset of those were selected that are both up-regulated compared to the pre-treatment condition in $\Delta efgA$ and compared to WT with formaldehyde (or down-regulated in both conditions). The analysis showed 64 annotated genes were up-regulated (Table 3.7). 10.9% of these genes belonged to chaperones and 4.69% of genes were related to DNA damage. There were 5 down-regulated genes in this category and 4 of them were involved with C₂, C₃ and C₄ compounds metabolism.

3.4 Discussion

Given the distinct patterns of OD increase despite viability loss for kanamycin stress, compared to a halt in OD but maintained viability for WT treated with formaldehyde, we can now relate whether gene expression tracks with OD or with viability. Cells treated with kanamycin showed exponential growth in terms of OD at 40 minutes despite the fact that cells had lost viability by 10-fold in terms of cell number. In terms of transcriptional changes, by 40 minutes cells showed nearly no response to kanamycin treatment. In contrast to kanamycin, formaldehyde showed a halt in viability at 40 minutes and at this time, the expression of many genes had already changed, with many as early as 5 minutes. In summary, cells appear to be unable to sense translation inhibition from kanamycin until they have already lost substantial viability. On the other hand, both gene expression and translation changed immediately when WT cells that have EfgA were treated with formaldehyde.

Although the temporal patterns of gene expression associated with translational inhibition were starkly different for kanamycin and formaldehyde, there was a large overlap in the effected genes. Almost half of the effected genes were common between the two treatments and the other half was specific to each (Figure 3.12). In addition, comparisons between specific functional groups (Table 3.3) showed most of these genes have been affected in both the formaldehyde and kanamycin responses. It is likely that most of these common expression changes represent the generic cellular consequences of translational arrest.

EfgA is clearly central to the cellular response to formaldehyde. This is seen in the timing and intensity of the gene expression change, where the WT response is stronger, and lasts longer than that seen for Δ efgA. Despite this, and the fact that growth of Δ efgA appears barely affected by formaldehyde, this strain still mounts a substantial gene expression response. Many genes are involved that provide hints as to the physiological problems directly caused by formaldehyde. Increased expression of genes such as chaperones suggests that protein misfolding was one of the key stresses encountered. Interestingly, comparison between the number of affected genes (Figure 3.10) and the specific functional groups (Table 3.3) both showed that this response to formaldehyde in Δ efgA is a subset of response to WT. This suggests that there are common physiological challenges faced, even in the absence of EfgA.

Genes encoding enzymes for formaldehyde metabolism showed changes upon the addition of exogenous formaldehyde, suggesting the cells actively regulate these activities to attempt to diminish the intracellular formaldehyde concentration. From the four *fdh* (formate dehydrogenase) gene clusters, two were up-regulated in both genotypes (*fdh2* and *fdh4*). For the other two clusters, *fdh1* was down-regulated in both genotypes, whereas *fdh3* was down-regulated in WT but the changes were not

significant in $\Delta efgA$. The *fhc* gene (formyltransferase/hydrolase complex) was also down-regulated in WT. The *mxg* genes (Ca-dependent methanol dehydrogenase) responsible for production of formaldehyde from methanol showed a significant down-regulation in both genotypes. This suggests that, although methanol oxidation was not the source of formaldehyde in this experiment, in the natural environment cells have the ability to sense intracellular formaldehyde and turn down expression of the enzymes that would be the typical source of this toxic intermediate.

Only a few of the beneficial loci observed during experimental evolution on elevated formaldehyde showed significant changes in expression upon exposure to formaldehyde. The *efgA* gene showed an up-regulation in WT, suggesting cells increase their ability to sense toxic formaldehyde. The *efgB* gene was significantly up-regulated in $\Delta efgA$ but not in WT. This finding showed that in the absence of EfgA, EfgB can sense formaldehyde or its consequences upon cellular physiology in a manner that does not occur in WT. The expression of neither the *marR*-like protein nor *def* (encoding PDF) changed significantly in any of the genotypes. In addition to *efgA* and *def*, there is one *efgA* homolog and one *def* homolog found in the *M. extorquens* PA1 genome. The *efgA* homolog showed a significant up-regulation in $\Delta efgA$ and a marginally significant ($p\text{-value}=5.01\times 10^{-2}$) up-regulation in WT, the *def* homolog showed a mild up-regulation in WT ($p\text{-value}=5.38\times 10^{-2}$) but there was no significant change observed in $\Delta efgA$. These data suggest that these uncharacterized homologs of *efgA* and *def* may play a role in formaldehyde stress, but further work is needed to test whether these apparent gene expression changes occur, and whether mutants lacking either gene have a formaldehyde-sensitivity phenotype.

The concerted decrease in cytochrome gene expression in response to the formaldehyde stress may be due to a similar response as is seen in the attenuation of respiration due to bacteriostatic antibiotics. The majority of bacteriostatic antibiotics have a role in translational inhibition, like kanamycin (Wilson, 2014), and these have been shown to decelerate respiration. On the other hand, bactericidal antibiotics accelerate respiration (Lobritz et al., 2015). Studies have shown that the effect of bactericidal antibiotics could be diminished when combined with bacteriostatic antibiotics (Brown and Alford, 1984; Crumplin and Smith, 1975; Deitz et al., 1966; Rocco and Overturf, 1982; Watanakunakorn and Guerriero, 1981; Weeks et al., 1981; Winslow et al., 1983). Translation inhibition may counteract the effect of stresses that accelerate respiration. The fact that bacteriostatic antibiotics could counteract the action of bactericidal antibiotics may suggest formaldehyde toxicity may work like bactericidal antibiotics. Given that the $\Delta efgA$ strain also showed some decrease in expression of respiratory proteins suggests this effect does not require EfgA; however, the strength of this response is stronger in WT.

One initially surprising response to formaldehyde stress was the down-regulation of flagellar components. Decreased synthesis of flagellar proteins has been studied in other examples of stresses, however. For example, in *Salmonella enterica*, a *ridA* mutant shows immotile phenotype despite being a metabolic gene. RidA has imine/enamine deaminase activity on 2-aminoacrylate, such that a *ridA* mutant accumulates toxic levels of 2-aminoacrylate. Consequently, transcriptomic analysis show genes involved in synthesis of flagellar assembly components were down-regulated (Borchert and Downs, 2017). *S. enterica* has shown to down-regulate flagellar components in nutritional stress. FlhD₄C₂ is an important regulator of flagellar synthesis. In poor media, Rflp protein (formerly Ydiv) showed up-regulation. In vitro observation showed Rflp binds FlhD₄C₂ and as a result inhibits FlhD₄C₂-dependent transcription of flagellar proteins (Wada et al., 2011). Moreover, biosynthesis of flagellar proteins could respond to cell envelope damage. RpoE and Rcs proteins can sense modification to cell envelope, and this process results in expression of Rflp. Rflp's action leads to degradation of FlhD₄C₂ via ClpXP protease and so down-regulation of flagella synthesis (Spöring et al., 2018).

ABC transporters were unique as a functional category with a substantial number of both up and down-regulated genes in response to formaldehyde. This family of transporters can have very different roles in eukaryotes and prokaryotes from importing nutrients, to exporting drugs and toxins (Davidson et al., 2008; Glavinas et al., 2004). The up-regulation of ABC transporters could indicate the cells attempting to export formaldehyde (or other adducts), or to decrease expression of transporters that allow the toxin into the cell.

Finally, perhaps the clearest link between cellular damage due to formaldehyde and its response to it is seen in the up-regulation of chaperones in both genotypes. As chaperones have role in response to misfolded proteins, up-regulation of chaperones indicates that cells are dealing with misfolded proteins caused by formaldehyde stress. Formaldehyde has been previously shown to have role in misfolding of proteins (He et al., 2010). This effect is incredibly wide-spread, even occurring in neuronal cells. Tau proteins stabilize microtubules, and formaldehyde stress leads to misfolding and amyloid-like aggregation of tau proteins (Nie et al., 2007). Correspondingly, formaldehyde stress has shown to induce up-regulation of heatshock proteins. Inhibition of hsp90 chaperone decreased human cell viability in otherwise non-toxic formaldehyde concentrations (Ortega-Atienza et al., 2016).

The results from this analysis emphasize how critical it was to have knowledge of the response to kanamycin in order to interpret the formaldehyde response. Almost half of the response observed for formaldehyde is actually shared between these two stressors. Although the shared response is quite interesting, this allowed focus upon the formaldehyde-specific aspects. Furthermore, comparing

responses between the two genotypes opens the window to the EfgA-independent responses. This analysis showed even though in $\Delta efgA$ mutant the response is attenuated compared to WT; but still the trend had similarities with the WT strain. According to the OD plots (Figure 3.2) $\Delta efgA$ still had growth in treatment with formaldehyde but expressing chaperones showed cells were dealing with consequences from formaldehyde toxicity and damaged proteins.

The fact that *efgB* showed an up-regulation in $\Delta efgA$ but not WT suggests EfgB could also be involved in sensing formaldehyde directly, or in sensing damages caused by formaldehyde. Moving forward, looking at $\Delta efgB$ and $\Delta efgA\Delta efgB$ could be the next step of the work. Part of the response in WT was shared with $\Delta efgA$ mutant, but given the fact that only $\Delta efgA$ showed up-regulation of *efgB*, assessing changes in $\Delta efgB$ could reveal the potential role of this protein in response to formaldehyde.

Investigating changes in frequent functional groups showed chaperones, cytochromes and ABC transporters were dominant. As we saw, down-regulation of cytochromes have been studied in translation inhibition response. This group and chaperones are likely to be upstream of translation inhibition, whereas flagellar proteins are likely to be the consequence of translation inhibition as there is not a clear link between motility and response to translation inhibition.

In addition to chaperones, DNA damage related proteins showed up-regulation and down-regulation. Formaldehyde is known as a potent agent to damage DNA (Grafstrom et al., 1983; Kawanishi et al., 2014) the fact that DNA damage related proteins were up-regulated suggests the cells activity to overcome the stress to DNA from formaldehyde toxicity.

Comparing significant up/down-regulated genes in $\Delta efgA$ treated with formaldehyde with WT in the same condition showed that, even though $\Delta efgA$ lacks the mechanism for translation inhibition, there were chaperones and DNA damage related proteins with relatively higher expressions compared to WT. This finding suggests in lack of EfgA (in $\Delta efgA$ mutant), where translation continues unabated, there were more damaged proteins and DNA. In addition, genes involved in metabolism of C₂, C₃ and C₄ compounds were down-regulated compared to WT. Why so much change in expression for genes encoding this particular part of metabolism? One hypothesis is that formaldehyde can interact directly with intermediate metabolites, creating damaged metabolites with adducts. These damaged metabolites could be harmful for the cells, or remove needed metabolites from the cell. Down-regulation of genes producing these metabolites could be a strategy to overcome the stress from metabolite damage. Alternatively, it could be that one or more enzymes in this part of metabolism are particularly sensitive to formaldehyde damage, and these gene expression changes are attempting to overcome these challenges.

Ultimately, it is important to note that all of the results here are based on RNA levels. Moving forward, mapping proteomic data (in progress) to current expression data brings another level to our understanding, as many of the reported genes in RNA-seq data are hypothetical proteins and knowing the protein profile will provide vital information about this translation inhibition system. In order to make the connection to functionality, transposon sequencing (Tn-seq) data could also add very important information to the current picture. Expression data provided information regarding candidate genes that have a role in translation inhibition, assessing the phenotype of mutant in these candidate genes could provide more information about the specific role of discussed candidate genes.

3.5 Supplementary material

Table 3.4 - P-values from t-tests of OD₆₀₀ in WT and Δ *efgA* in kanamycin (left) and formaldehyde (right) compared to the no-stressor in different timepoint, significant timepoints are shown in blue.

Time (minutes)	Kanamycin		Formaldehyde	
	WT	Δ <i>efgA</i>	WT	Δ <i>efgA</i>
-45	1.00E+00	1.00E+00	1.00E+00	1.00E+00
5	5.03E-01	4.77E-01	7.70E-01	7.83E-01
20	3.21E-01	9.42E-02	1.42E-01	7.67E-01
40	3.41E-01	5.61E-01	7.48E-03	2.49E-01
60	1.36E-01	5.55E-01	2.35E-03	2.87E-02
90	8.43E-01	8.21E-01	6.58E-02	1.94E-01
180	2.25E-01	1.11E-01	1.22E-03	4.49E-02
360	2.01E-03	1.67E-03	1.44E-04	9.07E-03
540	1.51E-04	2.43E-03	7.36E-04	1.10E-02

Table 3.5 - Up and down-regulated genes in WT with no-stressor compared to WT pre-treatment from 5 minutes to 180 minutes.

	Gene	Description
Up-regulated	Mext_0564	Secretion protein HlyD family protein
	Mext_0565	ABC transporter related
	Mext_0740	Chaperonin Cpn10
	Mext_1359	ABC transporter related
	Mext_1360	Cytochrome bd ubiquinol oxidase subunit I
	Mext_1361	Cytochrome d ubiquinol oxidase, subunit II
	Mext_2199	Hypothetical protein
	Mext_3120	Hypothetical protein
	Mext_3499	Putative transcriptional regulatory protein, Crp/Fnr family
	Mext_3500	UspA domain protein
	Mext_3502	Transport-associated
	Mext_3504	UspA domain protein
	Mext_3508	UspA domain protein
	Mext_3509	Metal-dependent phosphohydrolase HD sub domain
Down-regulated	Mext_0565	ABC transporter related
	Mext_0566	ABC-2 type transporter
	Mext_0567	Phosphoketolase
	Mext_1355	UspA domain protein
	Mext_1356	Cytochrome c class I
	Mext_1357	Cyclic nucleotide-binding
	Mext_1358	ABC transporter related
	Mext_1360	Cytochrome bd ubiquinol oxidase subunit I
	Mext_1361	Cytochrome d ubiquinol oxidase, subunit II
	Mext_1362	Cyd operon protein YbgT
	Mext_3498	Heat shock protein Hsp20
	Mext_3499	Putative transcriptional regulatory protein, Crp/Fnr family
	Mext_3500	UspA domain protein
	Mext_3502	Transport-associated
	Mext_3504	UspA domain protein
	Mext_3509	Metal-dependent phosphohydrolase HD sub domain

Table 3.6 - Up and down-regulated genes in $\Delta efgA$ with no-stressor compared to $\Delta efgA$ pre-treatment from 5 minutes to 180 minutes.

	Gene	Description
Up-regulated	Mext_0740	Chaperonin Cpn10
	Mext_2198	Hypothetical protein
	Mext_4782	Chaperonin GroEL
	Mext_4783	Chaperonin Cpn10
Down-regulated	Mext_0566	ABC-2 type transporter
	Mext_0567	Phosphoketolase
	Mext_1355	UspA domain protein
	Mext_1358	ABC transporter related
	Mext_1361	Cytochrome d ubiquinol oxidase, subunit II
	Mext_1362	Cyd operon protein YbgT

Table 3.7 - Genes that were up-regulated in $\Delta efgA$ at 5 minutes with formaldehyde compared to WT with formaldehyde at 5 minutes. The first column for each genotype shows the \log_2 FoldChange and the second column indicates the FDR adjusted p-value. These genes were also up-regulated compared to $\Delta efgA$ pre-treatment. Chaperones and heat shock proteins are colored in gray and DNA damage related proteins are colored in blue.

		$\Delta efgA$		WT	
Gene	Description	\log_2 FC	padj	\log_2 FC	padj
Mext_1801	Phosphoenolpyruvate carboxylase	3.081387	5.05E-49	0.997728	0.000356
Mext_1058	4-Diphosphocytidyl-2C-methyl-D-erythritol synthase	2.94557	4.51E-06	-1.21385	0.033793
Mext_3495	ABC transporter related	2.895307	2.43E-09	0.554093	0.352226
Mext_3819	Esterase, PHB depolymerase family	2.637244	5.38E-34	1.85177	1.27E-15
Mext_4556	Heat shock protein Hsp20	2.549191	3.61E-28	3.766948	1.23E-32
Mext_0646	Protein of unknown function DUF6 transmembrane	2.467646	1.04E-14	-0.61988	0.027858
Mext_3496	Acyl-CoA dehydrogenase type 2 domain	2.415107	5.64E-08	0.487878	0.248391
Mext_1802	Citrate (pro-3S)-lyase	2.39795	2.31E-24	0.306829	0.189875

Mext_2646	Nicotinate-nucleotide pyrophosphorylase	2.09155	1.44E-14	-0.11514	0.734087
Mext_2346	ATP-dependent chaperone ClpB	2.081596	4.73E-24	1.862724	3.94E-08
Mext_3411	Pyridine nucleotide-disulfide oxidoreductase family	2.067144	3.34E-08	0.389551	0.276443
Mext_2140	AMP-dependent synthetase and ligase	1.952313	1.63E-20	2.072552	7.37E-26
Mext_1798	Formiminotransferase- cyclodeaminase	1.914691	1.83E-15	1.157358	3.78E-08
Mext_2018	Protein of unknown function DUF477	1.888622	7.92E-09	-0.39952	0.180779
Mext_1800	Succinyl-CoA synthetase, alpha subunit	1.85894	4.13E-18	3.135273	3.24E-47
Mext_2252	Peptidase S16 lon domain protein	1.837365	1.90E-13	-0.70524	0.000183
Mext_4055	Excinuclease ABC, B subunit	1.816206	3.66E-15	0.19622	0.336486
Mext_4335	Cytochrome o ubiquinol oxidase, subunit III	1.699063	8.57E-16	-0.39653	0.128314
Mext_1796	D-isomer specific 2- hydroxyacid dehydrogenase NAD-binding	1.607575	1.43E-10	1.843014	1.50E-14
Mext_2019	Protein of unknown function DUF477	1.598352	1.04E-09	0.459044	0.055295
Mext_2418	ATP-dependent protease La	1.593469	3.40E-15	0.215091	0.61458
Mext_3410	Aliphatic sulfonates family ABC transporter, periplasmic ligand-binding protein	1.58769	2.90E-06	0.750335	0.010808
Mext_3931	Short-chain dehydrogenase/reductase SDR	1.55766	1.14E-09	-0.40911	0.078071
Mext_4352	Protein of unknown function DUF81	1.520888	4.60E-06	1.951716	4.73E-07

Mext_2380	Ribonuclease R	1.51174	3.69E-11	0.313949	0.146181
Mext_3979	FeS assembly protein SufD	1.510331	6.06E-08	-0.25744	0.257566
Mext_1797	Methylenetetrahydrofolate dehydrogenase (NADP(+))	1.502314	1.20E-12	1.893034	2.56E-21
Mext_1763	Bacterioferritin	1.450104	2.63E-07	0.643257	0.013747
Mext_1566	Integral membrane protein TerC	1.437209	2.03E-07	0.807545	0.00127
Mext_2961	Chaperone protein DnaJ	1.429271	2.96E-12	0.00959	0.965296
Mext_0645	Heat shock protein HslVU, ATPase subunit HslU	1.428986	5.65E-11	0.251611	0.201205
Mext_4508	Transporter, hydrophobe/amphiphile efflux-1 (HAE1) family	1.411288	3.47E-13	0.183119	0.323478
Mext_1178	Cysteine desulfurase, SufS subfamily	1.402077	1.49E-09	1.001465	1.35E-07
Mext_4406	Formate dehydrogenase, alpha subunit	1.377074	5.23E-11	2.255245	2.14E-32
Mext_2347	Metallophosphoesterase	1.359863	7.90E-08	-0.23669	0.3902
Mext_3978	FeS assembly ATPase SufC	1.309413	2.52E-09	-0.02909	0.920377
Mext_3670	Luciferase family protein	1.270375	1.63E-06	1.86665	4.76E-15
Mext_2960	Chaperone protein DnaK	1.249301	1.48E-08	1.586782	1.54E-14
Mext_3409	Aliphatic sulfonates family ABC transporter, periplasmic ligand-binding protein	1.22812	6.70E-06	1.699354	7.19E-13
Mext_3119	DNA topoisomerase IV, B subunit	1.204924	1.53E-08	-0.29243	0.144598
Mext_4336	Cytochrome o ubiquinol oxidase, subunit I	1.195474	6.86E-09	-0.19913	0.302381
Mext_2645	L-aspartate oxidase	1.188335	6.11E-07	1.127834	4.60E-05
Mext_1621	Protein of unknown function DUF1150	1.179329	4.67E-07	3.677721	2.74E-44
Mext_0606	Adenylyl cyclase class-3/4/guanylyl cyclase	1.169627	6.85E-07	0.168274	0.381391

Mext_3493	Inner-membrane translocator	1.121913	0.000766	1.509406	1.14E-07
Mext_2071	Coenzyme A transferase	1.113319	1.71E-08	1.685516	1.24E-26
Mext_1620	Heat shock protein Hsp20	1.090111	6.03E-06	2.651707	2.39E-30
Mext_4439	Double-strand break repair protein AddB	1.076873	2.49E-05	0.193613	0.44662
Mext_3977	FeS assembly protein SufB	1.057129	1.55E-06	0.746387	8.37E-05
Mext_3768	Putative transcriptional regulator, TetR family	1.033054	0.000288	0.86561	8.61E-06
Mext_1052	DNA topoisomerase IV, A subunit	1.025229	1.90E-06	0.203547	0.307411
Mext_2388	Methylmalonyl-CoA mutase	1.00834	5.27E-06	1.142207	2.39E-11
Mext_2515	Binding-protein-dependent transport systems inner membrane component	0.978845	4.25E-05	0.458072	0.007673
Mext_4782	Chaperonin GroEL	0.975936	0.000244	1.35235	0.001212
Mext_2254	Import inner membrane translocase subunit Tim44	0.956253	0.000269	1.825984	9.18E-08
Mext_1799	Succinyl-CoA synthetase, beta subunit	0.940191	4.34E-06	4.04627	7.73E-146
Mext_0660	D-3-phosphoglycerate dehydrogenase	0.929755	1.26E-05	1.141419	5.55E-12
Mext_0414	Formate--tetrahydrofolate ligase	0.902436	0.000305	2.149151	1.97E-46
Mext_3891	Excinuclease ABC, A subunit	0.891241	5.23E-05	0.620747	0.000217
Mext_0914	Aldehyde oxidase and xanthine dehydrogenase molybdopterin binding	0.828596	0.000107	2.221593	1.06E-35
Mext_2139	2-Dehydropantoate 2-reductase	0.815672	0.000133	3.21688	1.31E-68
Mext_2138	Fumarylacetoacetate (FAA) hydrolase	0.765552	0.000361	3.771497	4.66E-80
Mext_2105	Oxidoreductase alpha (molybdopterin) subunit	0.747819	0.000896	2.708724	1.62E-46

Table 3.8 - Genes that were down-regulated in $\Delta efgA$ at 5 minutes with formaldehyde, compared to WT with formaldehyde at 5 minutes. The first column for each genotype shows the \log_2 FoldChange and the second column indicates the FDR adjusted p-value. These genes were also down-regulated compared to the $\Delta efgA$ pre-treatment.

Gene	Description	$\Delta efgA$		WT	
		\log_2FC	padj	\log_2FC_{WT}	padj _{WT}
Mext_0854	Glycine cleavage system H protein	-1.34741	5.86E-05	-1.25403	1.51E-07
Mext_1509	Malate--quinone oxidoreductase	-1.11214	9.43E-06	-0.78313	9.76E-06
Mext_0853	Glycine dehydrogenase	-1.07712	4.86E-07	-1.17646	5.36E-14
Mext_1639	Phosphoenolpyruvate carboxykinase (ATP)	-1.06989	5.18E-05	-1.47383	7.22E-15
Mext_2790	Dihydrolipoamide dehydrogenase	-0.8917	7.80E-05	-0.79134	1.26E-05

4 Conclusion

Computational methods can inform biology by reducing complex dynamics to simple processes, testing mechanistic hypotheses, analyzing high-throughput data and finding patterns that are not obvious at first glance. In my thesis, I used two computational approaches to investigate the microbial stress response in a bacterial system where the stressor is normally produced intracellularly. By pairing empirical data with mathematical modeling and statistical analyses, I described phenotypic changes in a population, and then moved on to identify cellular components that may contribute to stress tolerance. Herein I presented my computational contributions toward understanding the physiological response of the model bacterium *M. extorquens* toward formaldehyde, a potentially lethal stress generated as an obligate intermediate of its own central metabolism. In the second chapter, I described how I used mathematical modeling to understand the mechanisms underlying the dynamics of tolerance distribution of an isogenic population. We saw growth and selection by death are not the only factors that shape changes in tolerance distribution of a population and that cells are able to change their phenotypic state (i.e., there is phenotypic movement between tolerance levels on the timescale of the experiment). Moreover, we saw that phenotypic movement depends on environmental conditions. This raises the possibility that cells might sense signals from their environment or their internal state, change their physiology in response to these signals, and consequently, their tolerance levels. To further understand mechanisms involved in the cellular response to a stressful environmental conditions, I used transcriptomic analysis in the second chapter. By exposing the wild-type and mutant strains to multiple stressors and tracking the gene expression profiles over time, I obtained a global picture of how cells respond to formaldehyde, found new patterns of specific stress response genes, and identified genes that potentially contribute to the tolerance distributions that I characterized in Chapter 2. This work provides significant groundwork for future work discovering and characterizing mechanisms underlying tolerance to formaldehyde.

Single-cell phenotypic heterogeneity in a performance trait like formaldehyde tolerance might emerge from expression heterogeneity in the gene expression level of a single gene, or perhaps requires heterogeneity in the expression of several genes. We therefore sought to identify the genes that may be involved in tolerance to formaldehyde with a transcriptomic analysis. We first needed to identify what sets of genes respond to the formaldehyde stress (Figure 4.1). My RNA-seq analysis showed a number of general stress response proteins such as chaperones, heatshock proteins and proteins involved in electron transport chain, like cytochromes, were differentially expressed when upon formaldehyde exposure. These genes could potentially have a role in tolerance at the subpopulation level. For example, having low cytochrome expression could confer high tolerance. Cytochromes have

a role in generating reactive oxygen (Liu et al., 2002) and down-regulation of these proteins could potentially decrease the chance of interaction between formaldehyde and reactive oxygen. Alternatively, the picture could be more complicated, such that tolerance might be the net effect of multiple upstream factors each with their own distribution of expression. For example, simultaneously having both high expression level of chaperones and low expression of cytochromes might be necessary in order to result in the high tolerance sub-population characterized in Chapter 2.

In the future, we can assess the expression of candidate genes responsible in tolerance at the single cell level. Unlike the eukaryotic systems, where single cell RNA-seq is trivial, bacterial systems face technical limitations for direct assays of RNA in single cells. This is due to their small volume, low mRNA amounts and short mRNA half lives (Gao et al., 2011; Wang et al., 2015). In bacterial systems, expression of genes at the single-cell level can be assessed through generation of transcriptional fusions to fluorescent proteins. Thus, investigating single-cell gene expression needs candidate genes, posing another limitation in relating gene expression to phenotype (e.g., tolerance). My RNA-seq results provided such candidate genes for these future studies aimed at describing how gene expression contributes to formaldehyde tolerance and phenotypic heterogeneity. For example, a fluorescent protein could be put under control of a cytochrome promoter, and fluorescence observed from this gene should generate an expression profile that matches RNA-seq data from bulk population experiments (i.e., expression should be lower upon formaldehyde exposure). If this were validated, then we could determine if the observed tolerant cells were biased toward a low expression level at the time of formaldehyde exposure. Using pairs of compatible fluorescent proteins, heterogeneity in multiple genes (e.g., a cytochrome and a chaperone) could be assessed simultaneously. Such an approach provides a framework to find the mechanisms underlying tolerance to formaldehyde.

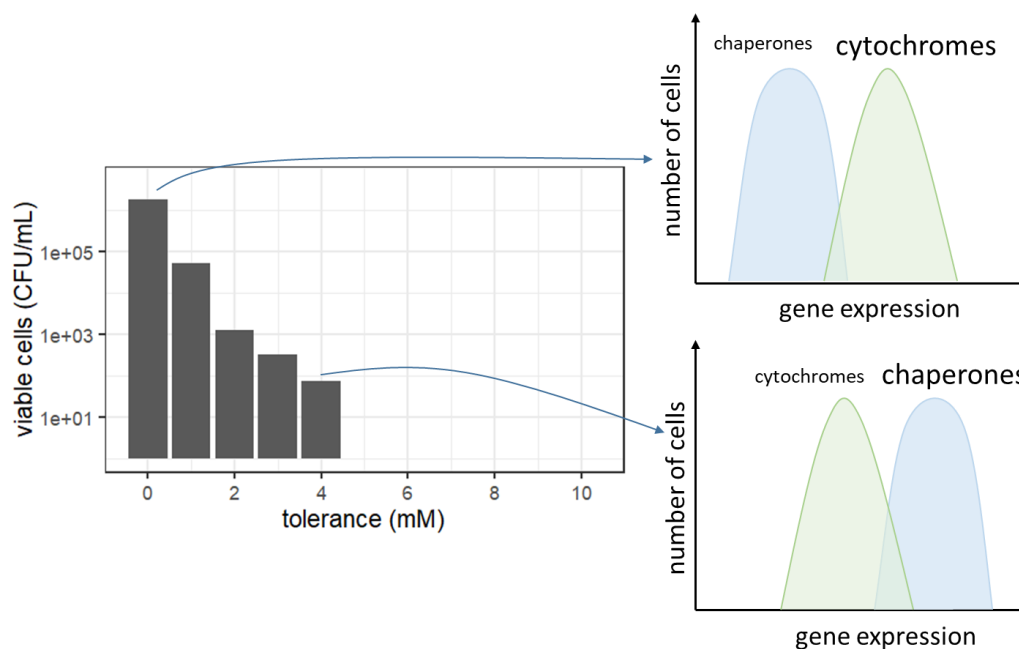


Figure 4.1- A phenotypic distribution of tolerance could be correlated with different gene expression profiles in each subpopulation. For example, different tolerance levels in *M. extorquens* may come from the combination of different expression levels of chaperones and cytochromes.

The fact that functional gene groups that responded to formaldehyde are not specific to *Methylobacterium* suggests that other bacteria might use similar mechanisms to tolerate formaldehyde stress. Formaldehyde is not used as a carbon source by most bacteria, but it is a by-product produced through various biochemical reactions in prokaryotes and eukaryotes (Roca et al., 2008). Accordingly, bacteria have a variety of mechanisms to detoxify formaldehyde at modest rates (Chen et al., 2016). Distributions of tolerance in other bacteria have shown that even non-methylotrophs such as *E. coli* are able to tolerate formaldehyde in low concentrations (Figure 4.2). My analyses showed that chaperones, heatshock proteins, cytochromes and ABC transporters are likely involved in managing formaldehyde stress in *M. extorquens*. It has been already shown that exposing formaldehyde to *Pseudomonas putida* induces expression of chaperones and DNA damage repair system (Roca et al., 2008). Performing RNA-seq analysis on other species of bacteria exposed to formaldehyde and comparing the results with our finding from *M. extorquens*, would be a straightforward way to find common mechanisms of tolerance to formaldehyde. Further, by knowing expression profile of genes involved in tolerance at the single-cell level (as discussed previously) we can potentially change tolerance level of other bacteria. Given that formaldehyde is considered to be an environmental hazard as well as a by-product of the manufacturing industries (Chen et al., 2016; Heck et al., 1990; Tang et al., 2009), there is great interest in finding ways to detoxify formaldehyde. Bacteria like *P. putida* can

consume environmental pollutant such as thiols (Marqués and Ramos, 1993). This bacterium is used in bioremediation of polluted environments and therefore could be a candidate organism to detoxify environments with formaldehyde pollution as well. For example, if further experimentation confirms that down-regulation of cytochromes has a demonstrable role in tolerance to formaldehyde in *M. extorquens*, down-regulation of cytochromes in *P. putida* might also increase its tolerance to formaldehyde.

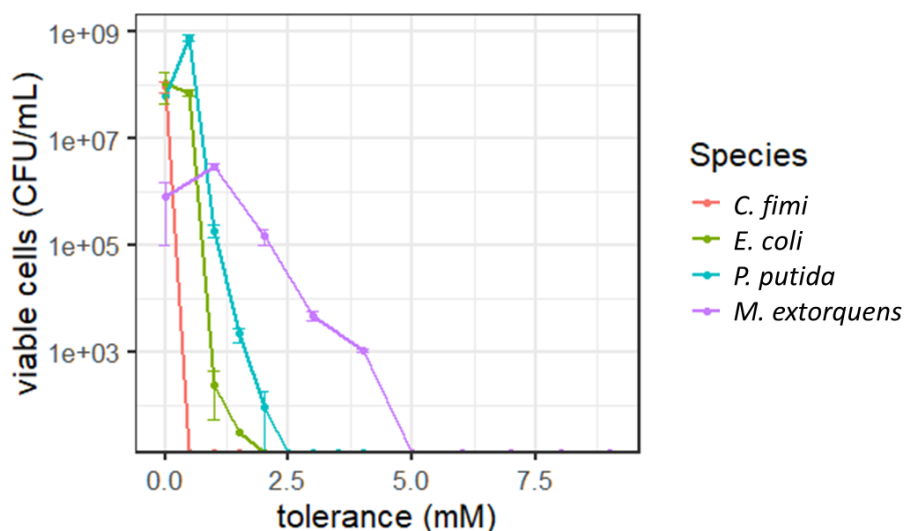


Figure 4.2 - Different tolerance distributions to formaldehyde for *Cellulomonas fimi*, *Escherichia coli*, *Pseudomonas putida*, or *Methylobacterium extorquens*.

Since the cellular response to formaldehyde suggests that generalized stress responses can contribute to tolerance, it raises the question: can tolerance level to the formaldehyde be indicative of the tolerance level to other stressors as well? It has been shown that tolerance to one stress can make cells more tolerant to other stressors, a phenomenon known as cross-stress protection (Dragosits et al., 2013). Chaperones, DNA repair system and ABC transporters all have roles in response to other stressors as well, so it is very likely that a cell tolerant to formaldehyde has already gained tolerance to at least some other stressors.

My model can provide a framework to investigate changes in potential distributions of other stressors. Growth and phenotypic movement terms could be applied to various bacteria or environmental stressors. My work has raised an interesting question regarding the behavior of death: does death from other stressors appear to be a sharp threshold between survival and loss of viability or are other types of death possible and dependent on the type of stressor? A threshold-based relationship of death as a function of tolerance could suggest there is a positive feedback loop in terms of protection and damage. For example, in formaldehyde treated cells, enzymes involved in formaldehyde oxidation that were

up-regulated are responsible for the detoxification of cells from excess formaldehyde. At lethal concentrations of formaldehyde, these enzymes may experience damage, which would result in a lowered capacity to overcome formaldehyde toxicity, thereby exacerbating loss of viability. A relative (non-threshold) relationship of death to tolerance levels would be expected if the system lacks such a positive feedback loop. In this case, formaldehyde toxicity would not directly damage the cell's machinery used to overcome toxicity, such that individuals with a tolerance level close to concentration of formaldehyde present do not experience a runaway loss of viability, and thus may exhibit an intermediate level of death.

As an extension to modeling only a single stressor, we could model a situation where bacteria face two stressors simultaneously in an environment. In this situation the number of cells with a given tolerance profile could be represented as $N(x_1, x_2, t)$ where x_1 represents the tolerance state of a cell for the first stressor and x_2 shows the tolerance state of the cell to the second stressor. This concept could be extended to $N(x_1, \dots, x_n, t)$ for any given number of stressors. To understand the response of cells in facing multiple stressors it is crucial to establish a relationship between different tolerance levels. Are tolerance levels linearly correlated with each other? Does tolerance to one stressor positively correlate with some stressors and negatively with some other stressors?

Phenotypic heterogeneity could be seen as an evolutionary capacity. Phenotypic heterogeneity may allow survival in stressful environment that would otherwise kill the average cell present, giving the population the ability to grow and have mutations arise that are selectively advantageous (Levin-Reisman et al., 2017). This phenomenon could be seen as a connection between Lamarckian and Darwinian evolution (Pisco et al., 2013). The fact that different bacterial species showed heterogeneity in tolerance to formaldehyde, including those that do not consume formaldehyde as a primary carbon source, suggests there could be an evolutionary advantage for a population to have a distribution of tolerance to a stressor.

This work described the response of *M. extorquens* to formaldehyde at the phenotypic level and investigated the transcriptome-phenotype relationship. Formaldehyde tolerance was found to involve genes that are part of the general stress response systems. This finding suggests that mechanisms in tolerance to formaldehyde could be common to other bacteria and to other stressors. Establishing the computational model of heterogeneity allowed for the investigation of processes involved in changing phenotype, helped us to understand the tolerance to formaldehyde in *M. extorquens* and provided a framework to study other stressors, as well as other bacterial species. Such investigations may have great importance in understanding the basic strategies used by organisms to overcome stress.

5 References

- Acar, M., Becskei, A., & van Oudenaarden, A. (2005). Enhancement of cellular memory by reducing stochastic transitions. *Nature*, *435*(7039), 228.
- Ackermann, M. (2015). A functional perspective on phenotypic heterogeneity in microorganisms. *Nature Reviews Microbiology*, *13*(8), 497.
- Adams, J. M. (1968). On the release of the formyl group from nascent protein. *Journal of Molecular Biology*, *33*(3), 571-589.
- Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716-723.
- Anders, S., & Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biology*, *11*(10), R106.
- Anders, S., Pyl, P. T., & Huber, W. (2015). HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics*, *31*(2), 166-169.
- Anthony, C., & Zatman, L. (1964). The microbial oxidation of methanol. 2. The methanol-oxidizing enzyme of *Pseudomonas* sp. M27. *Biochemical Journal*, *92*(3), 614.
- Artemova, T., Gerardin, Y., Dudley, C., Vega, N. M., & Gore, J. (2015). Isolated cell behavior drives the evolution of antibiotic resistance. *Molecular Systems Biology*, *11*(7), 822.
- Balaban, N. Q., Merrin, J., Chait, R., Kowalik, L., & Leibler, S. (2004). Bacterial persistence as a phenotypic switch. *Science*, *305*(5690), 1622-1625.
- Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. *arXiv preprint arXiv:1406.5823*.
- Bergmiller, T., & Ackermann, M. (2011). Pole age affects cell size and the timing of cell division in *Methylobacterium extorquens* AM1. *Journal of Bacteriology*, *193* (19) 5216-5221.
- Biggar, S. R., & Crabtree, G. R. (2001). Cell signaling can direct either binary or graded transcriptional responses. *The EMBO Journal*, *20*(12), 3167-3176.
- Bigger, J. (1944). Treatment of staphylococcal infections with penicillin by intermittent sterilisation. *The Lancet*, *244*(6320), 497-500.
- Bogel, G., Schrempf, H., & de Orué Lucana, D. O. (2009). The heme-binding protein HbpS regulates the activity of the *Streptomyces reticuli* iron-sensing histidine kinase SenS in a redox-dependent manner. *Amino Acids*, *37*(4), 681.
- Bojsen, R., Regenberg, B., Gresham, D., & Folkesson, A. (2016). A common mechanism involving the TORC1 pathway can lead to amphotericin B-persistence in biofilm and planktonic *Saccharomyces cerevisiae* populations. *Scientific Reports*, *6*, 21874.

- Borchert, A. J., & Downs, D. M. (2017). Endogenously generated 2-aminoacrylate inhibits motility in *Salmonella enterica*. *Scientific Reports*, 7(1), 12971.
- Brown, T., & Alford, R. (1984). Antagonism by chloramphenicol of broad-spectrum beta-lactam antibiotics against *Klebsiella pneumoniae*. *Antimicrobial Agents and Chemotherapy*, 25(4), 405-407.
- Chang, H. H., Hemberg, M., Barahona, M., Ingber, D. E., & Huang, S. (2008). Transcriptome-wide noise controls lineage choice in mammalian progenitor cells. *Nature*, 453(7194), 544.
- Chastanet, A., Vitkup, D., Yuan, G.-C., Norman, T. M., Liu, J. S., & Losick, R. M. (2010). Broadly heterogeneous activation of the master regulator for sporulation in *Bacillus subtilis*. *Proceedings of the National Academy of Sciences*, 107(18), 8486-8491.
- Chen, N. H., Djoko, K. Y., Veyrier, F. J., & McEwan, A. G. (2016). Formaldehyde stress responses in bacterial pathogens. *Frontiers in Microbiology*, 7, 257.
- Choi, P. J., Cai, L., Frieda, K., & Xie, X. S. (2008). A stochastic single-molecule event triggers phenotype switching of a bacterial cell. *Science*, 322(5900), 442-446.
- Crowther, G. J., Kosály, G., & Lidstrom, M. E. (2008). Formate as the main branch point for methylotrophic metabolism in *Methylobacterium extorquens* AM1. *Journal of Bacteriology*, 190(14), 5057-5062.
- Crumplin, G., & Smith, J. (1975). Nalidixic acid: an antibacterial paradox. *Antimicrobial Agents and Chemotherapy*, 8(3), 251-261.
- Davidson, A. L., Dassa, E., Orelle, C., & Chen, J. (2008). Structure, function, and evolution of bacterial ATP-binding cassette systems. *Microbiology and Molecular Biology Reviews*, 72(2), 317-364.
- de Orué Lucana, D. O., Bogel, G., Zou, P., & Groves, M. R. (2009). The oligomeric assembly of the novel haem-degrading protein HbpS is essential for interaction with its cognate two-component sensor kinase. *Journal of Molecular Biology*, 386(4), 1108-1122.
- Deitz, W. H., Cook, T. M., & Goss, W. A. (1966). Mechanism of action of nalidixic acid on *Escherichia coli* III. Conditions required for lethality. *Journal of Bacteriology*, 91(2), 768-773.
- Delaney, N. F., Kaczmarek, M. E., Ward, L. M., Swanson, P. K., Lee, M.-C., & Marx, C. J. (2013). Development of an optimized medium, strain and high-throughput culturing methods for *Methylobacterium extorquens*. *PloS One*, 8(4), e62957.
- Delaney, N. F., Rojas Echenique, J. I., & Marx, C. J. (2013). Clarity: an open-source manager for laboratory automation. *Journal of Laboratory Automation*, 18(2), 171-177.
- Deris, J. B., Kim, M., Zhang, Z., Okano, H., Hermsen, R., Groisman, A., & Hwa, T. (2013). The innate growth bistability and fitness landscapes of antibiotic-resistant bacteria. *Science*, 342(6162), 1237435.

- Dragosits, M., Mozhayskiy, V., Quinones-Soto, S., Park, J., & Tagkopoulos, I. (2013). Evolutionary potential, cross-stress behavior and the genetic basis of acquired stress resistance in *Escherichia coli*. *Molecular Systems Biology*, 9(1), 643.
- Elowitz, M. B., Levine, A. J., Siggia, E. D., & Swain, P. S. (2002). Stochastic gene expression in a single cell. *Science*, 297(5584), 1183-1186.
- Ferrell Jr, J. E. (2012). Bistability, bifurcations, and Waddington's epigenetic landscape. *Current Biology*, 22(11), R458-R466.
- Gao, W., Zhang, W., & Meldrum, D. R. (2011). RT-qPCR based quantitative analysis of gene expression in single bacterial cells. *Journal of Microbiological Methods*, 85(3), 221-227.
- Gillet, J.-P., & Gottesman, M. M. (2010). Mechanisms of multidrug resistance in cancer. *Multi-Drug Resistance in Cancer* (pp. 47-76): Springer.
- Glavinas, H., Krajcsi, P., Cserepes, J., & Sarkadi, B. (2004). The role of ABC transporters in drug resistance, metabolism and toxicity. *Current Drug Delivery*, 1(1), 27-42.
- Grafstrom, R. C., Fornace, A. J., Autrup, H., Lechner, J. F., & Harris, C. C. (1983). Formaldehyde damage to DNA and inhibition of DNA repair in human bronchial cells. *Science*, 220(4593), 216-218.
- Hasty, J., Pradines, J., Dolnik, M., & Collins, J. J. (2000). Noise-based switches and amplifiers for gene expression. *Proceedings of the National Academy of Sciences*, 97(5), 2075-2080.
- He, R., Lu, J., & Miao, J. (2010). Formaldehyde stress. *Science China Life Sciences*, 53(12), 1399-1404.
- Heck, d. H. A., Casanova, M., & Starr, T. B. (1990). Formaldehyde toxicity—new understanding. *Critical Reviews in Toxicology*, 20(6), 397-426.
- Hindmarsh, A. C. (1983). ODEPACK, a systematized collection of ODE solvers. *Scientific Computing*, 55-64.
- Huh, D., & Paulsson, J. (2011). Non-genetic heterogeneity from stochastic partitioning at cell division. *Nature Genetics*, 43(2), 95.
- Igoshin, O. A., Alves, R., & Savageau, M. A. (2008). Hysteretic and graded responses in bacterial two-component signal transduction. *Molecular Microbiology*, 68(5), 1196-1215.
- Johnson, N. L. (1949). Systems of frequency curves generated by methods of translation. *Biometrika*, 36(1/2), 149-176.
- Kawanishi, M., Matsuda, T., & Yagi, T. (2014). Genotoxicity of formaldehyde: molecular basis of DNA damage and mutation. *Frontiers in Environmental Science*, 2, 36.
- Keijser, B. J., Ter Beek, A., Rauwerda, H., Schuren, F., Montijn, R., van der Spek, H., & Brul, S. (2007). Analysis of temporal gene expression during *Bacillus subtilis* spore germination and outgrowth. *Journal of Bacteriology*, 189(9), 3624-3634.

- Kim, S., Cho, Y.-J., Song, E.-S., Lee, S. H., Kim, J.-G., & Kang, L.-W. (2016). Time-resolved pathogenic gene expression analysis of the plant pathogen *Xanthomonas oryzae* pv. *oryzae*. *BMC Genomics*, *17*(1), 345.
- Kiviet, D. J., Nghe, P., Walker, N., Boulineau, S., Sunderlikova, V., & Tans, S. J. (2014). Stochasticity of metabolism and growth at the single-cell level. *Nature*, *514*(7522), 376.
- Kohanski, M. A., Dwyer, D. J., & Collins, J. J. (2010). How antibiotics kill bacteria: from targets to networks. *Nature Reviews Microbiology*, *8*(6), 423.
- Korch, S. B., Henderson, T. A., & Hill, T. M. (2003). Characterization of the hipA7 allele of *Escherichia coli* and evidence that high persistence is governed by (p) ppGpp synthesis. *Molecular Microbiology*, *50*(4), 1199-1213.
- Kussell, E., Kishony, R., Balaban, N. Q., & Leibler, S. (2005). Bacterial persistence a model of survival in changing environments. *Genetics*, *169*(4), 1807-1814.
- LaFleur, M. D., Kumamoto, C. A., & Lewis, K. (2006). *Candida albicans* biofilms produce antifungal-tolerant persister cells. *Antimicrobial Agents and Chemotherapy*, *50*(11), 3839-3846.
- Levin-Reisman, I., Ronin, I., Gefen, O., Braniss, I., Shoshitashvili, N., & Balaban, N. Q. (2017). Antibiotic tolerance facilitates the evolution of resistance. *Science*, eaaj2191.
- Lindner, A. B., Madden, R., Demarez, A., Stewart, E. J., & Taddei, F. (2008). Asymmetric segregation of protein aggregates is associated with cellular aging and rejuvenation. *Proceedings of the National Academy of Sciences*, *105*(8), 3076-3081.
- Liu, Y., Fiskum, G., & Schubert, D. (2002). Generation of reactive oxygen species by the mitochondrial electron transport chain. *Journal of Neurochemistry*, *80*(5), 780-787.
- Lobritz, M. A., Belenky, P., Porter, C. B., Gutierrez, A., Yang, J. H., Schwarz, E. G., Collins, J. J. (2015). Antibiotic efficacy is linked to bacterial cellular respiration. *Proceedings of the National Academy of Sciences*, 201509743.
- Lorenzi, T., Chisholm, R. H., & Clairambault, J. (2016). Tracking the evolution of cancer cell populations through the mathematical lens of phenotype-structured equations. *Biology Direct*, *11*(1), 43.
- Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, *15*(12), 550.
- Maamar, H., & Dubnau, D. (2005). Bistability in the *Bacillus subtilis* K-state (competence) system requires a positive feedback loop. *Molecular Microbiology*, *56*(3), 615-624.
- Maamar, H., Raj, A., & Dubnau, D. (2007). Noise in gene expression determines cell fate in *Bacillus subtilis*. *Science*, *317*(5837), 526-529.
- Marqués, S., & Ramos, J. L. (1993). Transcriptional control of the *Pseudomonas putida* TOL plasmid catabolic pathways. *Molecular Microbiology*, *9*(5), 923-929.

- Marx, C. J., Chistoserdova, L., & Lidstrom, M. E. (2003). Formaldehyde-detoxifying role of the tetrahydromethanopterin-linked pathway in *Methylobacterium extorquens* AM1. *Journal of Bacteriology*, 185(24), 7160-7168.
- Marx, C. J., Van Dien, S. J., & Lidstrom, M. E. (2005). Flux analysis uncovers key role of functional redundancy in formaldehyde metabolism. *PLoS Biology*, 3(2), e16.
- McAdams, H. H., & Arkin, A. (1997). Stochastic mechanisms in gene expression. *Proceedings of the National Academy of Sciences*, 94(3), 814-819.
- McALLISTER, C. F., & Lepo, J. (1983). Succinate transport by free-living forms of *Rhizobium japonicum*. *Journal of Bacteriology*, 153(3), 1155-1162.
- Merow, C., Dahlgren, J. P., Metcalf, C. J. E., Childs, D. Z., Evans, M. E., Jongejans, E., McMahon, S. M. (2014). Advancing population ecology with integral projection models: a practical guide. *Methods in Ecology and Evolution*, 5(2), 99-110.
- Nelder, J. A., & Mead, R. (1965). A simplex method for function minimization. *The Computer Journal*, 7(4), 308-313.
- Nie, C. L., Wang, X. S., Liu, Y., Perrett, S., & He, R. Q. (2007). Amyloid-like aggregates of neuronal tau induced by formaldehyde promote apoptosis of neuronal cells. *BMC Neuroscience*, 8(1), 9.
- Ortega-Atienza, S., Rubis, B., McCarthy, C., & Zhitkovich, A. (2016). Formaldehyde is a potent Proteotoxic stressor causing rapid heat shock transcription factor 1 activation and Lys48-linked Polyubiquitination of proteins. *The American Journal of Pathology*, 186(11), 2857-2868.
- Ortiz de Orué Lucana, D., Hickey, N., Hensel, M., Klare, J. P., Geremia, S., Tiufiakova, T., & Torda, A. E. (2016). The Crystal Structure of the C-Terminal Domain of the *Salmonella enterica* PduO Protein: An Old Fold with a New Heme-Binding Mode. *Frontiers in Microbiology*, 7, 1010.
- Perthame, B. (2015). Parabolic equations in biology. *Parabolic Equations in Biology* (pp. 1-21): Springer.
- Pinto, D., Santos, M. A., & Chambel, L. (2015). Thirty years of viable but nonculturable state research: unsolved molecular mechanisms. *Critical Reviews in Microbiology*, 41(1), 61-76.
- Pisco, A. O., Brock, A., Zhou, J., Moor, A., Mojtahedi, M., Jackson, D., & Huang, S. (2013). Non-Darwinian dynamics in therapy-induced cancer drug resistance. *Nature Communications*, 4, 2467.
- Plahte, E., Mestl, T., & Omholt, S. W. (1995). Feedback loops, stability and multistationarity in dynamical systems. *Journal of Biological Systems*, 3(02), 409-413.
- Raj, A., & van Oudenaarden, A. (2008). Nature, nurture, or chance: stochastic gene expression and its consequences. *Cell*, 135(2), 216-226.

- Reams, A. B., Kofoed, E., Kugelberg, E., & Roth, J. R. (2012). Multiple pathways of duplication formation with and without recombination (RecA) in *Salmonella enterica*. *Genetics*, 112.142570.
- Roca, A., Rodríguez-Herva, J. J., Duque, E., & Ramos, J. L. (2008). Physiological responses of *Pseudomonas putida* to formaldehyde during detoxification. *Microbial Biotechnology*, 1(2), 158-169.
- Rocco, V., & Overturf, G. (1982). Chloramphenicol inhibition of the bactericidal effect of ampicillin against *Haemophilus influenzae*. *Antimicrobial Agents and Chemotherapy*, 21(2), 349-351.
- Savageau, M. (1999). Design of gene circuitry by natural selection: analysis of the lactose catabolic system in *Escherichia coli*. *Portland Press Limited*, 264-270.
- Smits, W. K., Eschevins, C. C., Susanna, K. A., Bron, S., Kuipers, O. P., & Hamoen, L. W. (2005). Stripping Bacillus: ComK auto-stimulation is responsible for the bistable response in competence development. *Molecular Microbiology*, 56(3), 604-614.
- Snoussi, E. H. (1998). Necessary conditions for multistationarity and stable periodicity. *Journal of Biological Systems*, 6(01), 3-9.
- Soetaert, K., & Meysman, F. (2012). Reactive transport in aquatic ecosystems: Rapid model prototyping in the open source software R. *Environmental Modelling & Software*, 32, 49-60.
- Soetaert, K., Petzoldt, T., & Setzer, R. W. (2010). Solving differential equations in R: package deSolve. *Journal of Statistical Software*, 33.
- Spörling, I., Felgner, S., Preuße, M., Eckweiler, D., Rohde, M., Häussler, S., Erhardt, M. (2018). Regulation of Flagellum Biosynthesis in Response to Cell Envelope Stress in *Salmonella enterica* Serovar Typhimurium. *mBio*, 9(3), e00736-00717.
- Süel, G. M., Garcia-Ojalvo, J., Liberman, L. M., & Elowitz, M. B. (2006). An excitable gene regulatory circuit induces transient cellular differentiation. *Nature*, 440(7083), 545.
- Süel, G. M., Kulkarni, R. P., Dworkin, J., Garcia-Ojalvo, J., & Elowitz, M. B. (2007). Tunability and noise dependence in differentiation dynamics. *Science*, 315(5819), 1716-1719.
- Tang, X., Bai, Y., Duong, A., Smith, M. T., Li, L., & Zhang, L. (2009). Formaldehyde in China: production, consumption, exposure levels, and health effects. *Environment International*, 35(8), 1210-1224.
- Thomas, R. (1981). On the relation between the logical structure of systems and their ability to generate multiple steady states or sustained oscillations. *Numerical Methods in the Study of Critical Phenomena* (pp. 180-193): Springer.
- Udekwi, K. I., Parrish, N., Ankomah, P., Baquero, F., & Levin, B. R. (2009). Functional relationship between bacterial cell density and the efficacy of antibiotics. *Journal of Antimicrobial Chemotherapy*, 63(4), 745-757.
- Van den Bergh, B., Fauvart, M., & Michiels, J. (2017). Formation, physiology, ecology, evolution and clinical importance of bacterial persisters. *FEMS Microbiology Reviews*, 41(3), 219-251.

- Veening, J.-W., Stewart, E. J., Berngruber, T. W., Taddei, F., Kuipers, O. P., & Hamoen, L. W. (2008). Bet-hedging and epigenetic inheritance in bacterial cell development. *Proceedings of the National Academy of Sciences*, *105*(11), 4393-4398.
- Vorholt, J. A. (2012). Microbial life in the phyllosphere. *Nature Reviews Microbiology*, *10*(12), 828-840.
- Vorholt, J. A., Marx, C. J., Lidstrom, M. E., & Thauer, R. K. (2000). Novel formaldehyde-activating enzyme in *Methylobacterium extorquens* AM1 required for growth on methanol. *Journal of Bacteriology*, *182*(23), 6645-6650.
- Wada, T., Morizane, T., Abo, T., Tominaga, A., Inoue-Tanaka, K., & Kutsukake, K. (2011). An EAL-domain protein YdiV acts as an anti-FlhD4C2 factor responsible for nutritional control of the flagellar regulon in *Salmonella enterica* serovar Typhimurium. *Journal of Bacteriology*, *193*(7) 1600-1611.
- Wang, J., Chen, L., Chen, Z., & Zhang, W. (2015). RNA-seq based transcriptomic analysis of single bacterial cells. *Integrative Biology*, *7*(11), 1466-1476.
- Watanakunakorn, C., & Guerriero, J. C. (1981). Interaction between vancomycin and rifampin against *Staphylococcus aureus*. *Antimicrobial Agents and Chemotherapy*, *19*(6), 1089.
- Weeks, J. L., Mason, E., & Baker, C. J. (1981). Antagonism of ampicillin and chloramphenicol for meningeal isolates of group B streptococci. *Antimicrobial Agents and Chemotherapy*, *20*(3), 281-285.
- Weinberger, L. S., Burnett, J. C., Toettcher, J. E., Arkin, A. P., & Schaffer, D. V. (2005). Stochastic gene expression in a lentiviral positive-feedback loop: HIV-1 Tat fluctuations drive phenotypic diversity. *Cell*, *122*(2), 169-182.
- Wilson, D. N. (2014). Ribosome-targeting antibiotics and mechanisms of bacterial resistance. *Nature Reviews Microbiology*, *12*(1), 35.
- Winslow, D., Damme, J., & Dieckman, E. (1983). Delayed bactericidal activity of beta-lactam antibiotics against *Listeria monocytogenes*: antagonism of chloramphenicol and rifampin. *Antimicrobial Agents and Chemotherapy*, *23*(4), 555-558.
- Zhi, J., Nightingale, C. H., & Quintiliani, R. (1986). A pharmacodynamic model for the activity of antibiotics against microorganisms under nonsaturable conditions. *Journal of Pharmaceutical Sciences*, *75*(11), 1063-1067.

6 Appendix

The following data files for chapter 3 can be found online in the “Supplemental files” section of ProQuest website. In each data file, the first and second columns show genes and their descriptions, the third and fourth columns show the \log_2 FoldChange and the FDR adjusted p-values respectively.

efgA_formaldehyde_5.csv: Up and down-regulated genes in Δ efgA with formaldehyde treatment at 5 minutes.

efgA_formaldehyde_20.csv: Up and down-regulated genes in Δ efgA with formaldehyde treatment at 20 minutes.

efgA_kanamycin_180.csv: Up and down-regulated genes in Δ efgA with kanamycin treatment at 180 minutes.

efgA_kanamycin_360.csv: Up and down-regulated genes in Δ efgA with kanamycin treatment at 360 minutes.

efgA_noStressor_360.csv: Up and down-regulated genes in Δ efgA with no-stressor treatment at 360 minutes.

WT_formaldehyde_5.csv: Up and down-regulated genes in WT with formaldehyde treatment at 5 minutes.

WT_formaldehyde_20.csv: Up and down-regulated genes in WT with formaldehyde treatment at 20 minutes.

WT_kanamycin_180.csv: Up and down-regulated genes in WT with kanamycin treatment at 180 minutes.

WT_kanamycin_360.csv: Up and down-regulated genes in WT with kanamycin treatment at 360 minutes.

WT_noStressor_360.csv: Up and down-regulated genes in WT with no-stressor treatment at 180 minutes.