REGULATORY ORGANIZATION AND TRANSCRIPTIONAL RESPONSE OF SPHINGOBIUM CHLOROPHENOLICUM TO THE ANTHROPOGENIC PESTICIDE PENTACHLOROPHENOL

by

JOE ROKICKI

B.S.E, Princeton University, 2008

A thesis submitted to the

Faculty of the Graduate School of the

University of Colorado in partial fulfillment

of the requirement for the degree of

Doctor of Philosophy

Department of Molecular, Cellular, and Developmental Biology

2015

This thesis entitled: Regulatory organization and transcriptional response of *Sphingobium chlorophenolicum* to the anthropogenic pesticide pentachlorophenol, written by Joe Rokicki, has been approved by the department of Molecular, Cellular, and Developmental Biology

---

Robin Dowell

Corrie Detweiler

Date: _____

The final copy of this thesis has been examined by the signatories and will find that the content and the form meet acceptable presentation standards of scholarly work in the above mentioned discipline.

Abstract

Rokicki, Joe Franklin (Ph.D., Molecular, Cellular, and Developmental Biology)

Regulatory Organization and Transcriptional Response of *Sphingobium chlorophenolicum* to the Anthropogenic Pesticide Pentachlorophenol

Thesis directed by Assistant Professor Robin Dowell

The sudden and widespread introduction of the pesticide pentachlorophenol (PCP) into the environment from 1930 to 1980 created a new global selection pressure on microbes. The subsequent isolation of a pentachlorophenol degrading bacterium, *S. chlorophenolicum*, provided a unique opportunity to study an early evolutionary response to the new selective pressure.

The minimal enzymatic pathway required to degrade PCP was laboriously determined before high throughput sequencing was possible through a highly targeted approach that proved effective but left many of the evolutionary questions that motivated the study of this pathway and this organism unanswered. Where did these genes come from? Did they originate from horizontal gene transfer, duplication and divergence, or recruitment? What are the regulatory mechanisms of this pathway? Are other genes induced by PCP? To answer these questions, a global perspective of the genome and transcriptome of the organism is required.

In this thesis, I ascertain and discuss the complete genome sequence of *S. chlorophenolicum*, I discuss the development of a bioinformatics tool to facilitate massively comparative microbial genomics, I uncover and examine the global transcriptional response of *S. chlorophenolicum* to PCP, and I take a detailed molecular look at the key transcription factors governing this regulatory network.

Dedication

My parents, Cathy and Jahn Rokicki, have supported me in every way possible from the minute I was born. I have no doubts that the type of scientist I aspire to be grew out of the curiosity and values they have modeled and encouraged for me my entire life. This dissertation is dedicated to them.

Acknowledgements

All my friends and peers at MCDB have created an incredible environment of scientific and emotional support. I would especially like to acknowledge Tim Read, Sam O'Hara and Rakel Salamander for countless conversations about science both indoors, with dry erase markers, and outdoors, hiking in the mountains and drinking out of CamelBaks.

I am indebted to Dylan Taatjes and Robin Dowell and their graduate students. They opened up their labs to me and made many of the experiments in this dissertation possible. All of the radioactivity work in this thesis was performed because of Dylan's generosity. Similarly, all the ChIP was performed with reagents and advice from the Dowell Lab.

I would like to thank Shelley for introducing me to these ideas and supporting me while I tried to get to the bottom of them.

Last but not least, I would like to thank my dog Zoey for getting me out of the lab and taking me on a walk from time to time.

Table of Contents

Table of Figures

Chapter 1 Introduction and Background

Introduction

PCP is a toxic pesticide that was introduced into the environment in vast quantities from the 1930s through the 1980s. In this section, I discuss the basis of the broad spectrum toxicity of the PCP molecule. I describe several organisms from diverse bacterial phyla, as well as a eukaryote, that are able to catalyze the complete or partial degradation of PCP. Finally, I discuss the PCP degradation pathway of *S. chlorophenolicum*, including the isolation of this organism and the techniques used to uncover the minimal enzymatic pathway it utilizes to degrade PCP.

The Introduction of Pentachlorophenol Into the Environment

Pentachlorophenol (PCP) became a popular broad-spectrum pesticide in the 1930's. PCP was used in the United States primarily as a wood preservative for telephone poles and railroad ties, extending their functional lifetimes by many years (Union, Pure, and Chemistry 1987; Cirelli 1978). The EPA designated PCP a controlled substance in 1978 at which point it was being produced worldwide in quantities of over 50 million kg / year. In 1987, PCP was banned by the EPA due to concerns of its widespread persistence in the environment and its status as a suspected carcinogen and teratogen (McAllister, Lee, and Trevors 1996).

The molecular structure of PCP is the basis for its toxicity. PCP consists of a benzene ring with one hydroxyl group and five chlorines. The strong electronegativity of the chlorines distributes the electron resonance of the phenol to such a degree that PCP is able to pass easily

through hydrophobic membranes both while it is neutral or negatively charged. These properties allow it to shuttle protons through a membrane, collapsing the proton gradient across it. Because all domains of life utilize proton gradients as a mechanism of energy generation, PCP has indiscriminate toxicity.

This ability to uncouple the proton gradient generated by oxidative phosphorylation from the energy generation of ATP synthase is a property of a class of molecules called, "uncouplers". In addition to PCP, many other molecules have been identified with this property. In bacteria, assault by phage can in some cases collapse the proton gradient. A stress response to uncoupling called phage shock response is conserved in many lineages of bacteria (Darwin 2005). Another famous uncoupler is dinitrophenol (DNP). It was used extensively after its discovery in the 1930s as a dieting drug (Cutting, Mehrtens, and Tainter 1933). Similar to PCP, DNP acts as an ionophore resulting in a nonproductive release of the energy from calories consumed. This nonproductive release of energy results in the production of heat. Physiologically, this property of heat generation has been harnessed evolutionarily in humans by the uncoupling protein thermogenin (UPC1) (Ricquier 1999). Thermogenin leaks protons across the membrane and generates heat in the process to drive non-shivering thermogenesis. This thermogenesis is the primary mechanism of heat generation in infants (Zaninovich et al. 2002; Blumberg, Deaver, and Kirby 1999).

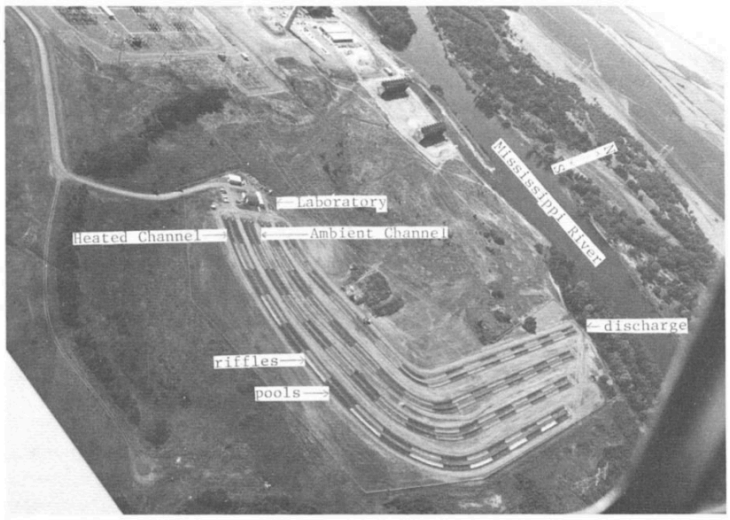
There are many examples in nature of evolutionary responses to proton gradient collapse. The degradation pathway of PCP in the bacterium *S. chlorophenolicum* is unique for several reasons. First, the degradation pathway is highly specific to PCP despite there being no known natural sources of PCP. Second, the evolutionary response to PCP stress is a relatively

elaborate pathway involving many enzymes and a complicated regulatory system. Compared to evolutionary adaptation that simply detoxify a compound in one step or exclude a compound from the cell, the PCP degradation pathway in *S. chlorophenolicum* a fascinating model evolutionary response.

Discovery and Isolation of *S. chlorophenolicum* L-1

To study the fate of PCP after it is introduced into the environment, Pignatello et al. dosed four artificial rivers at the Monticello Ecological Research Station with PCP concentrations ranging from 0 to 432 ug/L continuously over the course of 16 weeks. The four 520-meter long artificial rivers were fed with water diverted from the Mississippi (see Figure 1.1) populating them with the natural microbial populations of the river. The researchers measured aqueous and sedimentary PCP concentrations, and sampled the microbial populations of the sediments and water throughout the experiment.

Figure 1.1 - Aerial view of the Monticello Ecological Research Center



This photo adapted from Nordlie et al. (Nordlie and Arthur 1981)

The researchers found that at the river surface PCP was rapidly degraded via photolysis, but this effect deteriorated rapidly as water depth increased and ultraviolet light penetration decreased. At the river bottom, PCP rapidly accumulated and persisted within the river sediments. Within 20 days, the accumulated PCP began to rapidly disappear. The phenomenon was attributed to microbial degradation and quickly surpassed photolysis as the dominant process of PCP degradation. Interestingly, total bacteria counts in the sediments were not adversely affected at any PCP concentration tested (Pignatello et al. 1983).

Saber et al. isolated and characterized the PCP degrading organisms from the river sediments of the research station, as well as from four geographically distinct soil samples from PCP contaminated areas (Saber and Crawford 1985). In every case, all PCP degrading bacterial strains isolated from both the river and soil samples proved to be members of the Sphingomonadaceae family. The champion PCP degrading strain from this experiment was deposited in the ATCC strain repository under the name *S. chlorophenolicum* L-1 and strain collection number 39273. Unspecified problems with strain propagation at ATCC resulted in this strain losing the ability to degrade PCP. The strain was resubmitted under the new strain collection number 52874.

Other Pentachlorophenol Degrading Organisms

Though *S. chlorophenolicum* is the best studied of the pentachlorophenol degrading organisms, it is not the only organism identified with this ability. Pentachlorophenol degradation has been found in many other Sphingomonads, as well as in several other phyla of

bacteria and in at least one eukaryote. Below, I describe examples of PCP degradation from across the tree of life.

Many members of the bacterial class Alphaproteobacteria, including *S. chlorophenolicum* L1 and several other strains of *Sphingobium and Novosphingobium,* have shown PCP degradation activity (Nohynek et al. 1995; Edgehill and Finn 1983). These strains were isolated from contaminated sites all over the world and show various levels of divergence. All Alphaproteobacteria that degrade PCP are Sphingomonads and contain a *pcpB* homolog (the first enzyme involved in the degradation pathway of PCP) as shown by PCR or genomic DNA hybridization assays.

Tiirola et al. made a case for the horizontal gene transfer of the *pcpB* gene among Sphingomonads isolated from a contaminated water source in Finland. By comparing the 16S ribosomal sequences and the *pcpB* gene sequences of 11 PCP degrading species isolated from groundwater in Finland, they found that the *pcpB* gene appeared unusually conserved among many of the Sphingomonads (Tiirola et al. 2002). The authors noted that despite examining other polychlorophenol degrading bacteria from the same environment, *pcpB* could only be detected in Sphingomonads, perhaps implying a taxonomic barrier confining the *pcpB* gene to that group. However, this observation could also be consistent with a pattern of vertical descent for *pcpB* in the Sphingomonads. Without genome sequences, it was impossible to look for other markers of horizontal gene transfer to support either conclusion.

In addition to the Alphaproteobacteria, *Desulfomonile tiedjei* DCB-1 (ATCC 49306) of the class Deltaproteobacteria, partially degrades PCP. *D. tiedjei* is an obligate anaerobe isolated from sewage sludge as part of a methanogenic consortium with the ability to degrade 3-

chlorobenzoic acid. Experiments have shown that *D. tiedjei* is able to partially degrade PCP to 2,4,6 trichlorophenol but only in the presence of 3-chlorobenzoic acid, likely acting as an inducer (Mohn and Kennedy 1992; Shelton and Tiedje 1984). Limiting oxygen or reductant resulted in the appearance of a tetrachlorophenol intermediate suggesting the reaction occurs in two steps, potentially involving multiple enzymes. In all conditions tested, degradation did not proceed past trichlorophenol.

A member of the class Gammaproteobacteria isolated from paper mill sludge, *Pseudomonas stutzeri CL7,* has also been shown to degrade PCP (Karn, Chakrabarty, and Reddy 2010). PCP degradation by this organism is accompanied by a stoichiometric release of chloride ions indicating complete mineralization of PCP.

Pentachlorophenol degradation is not confined to the Proteobacteria phylum. A member of the Actinobacter phylum, *Mycobacterium chlorophenolicum,* isolated from PCP contaminated lake sediment in Finland, was shown to completely degrade PCP (Briglia et al. 1994; Apajalahti and Salkinoja-Salonen 1986; Crawford, Jung, and Strap 2007). Recent genome sequencing revealed the presence of a distant *pcpB* homolog in this organism, although it is unclear if this enzyme is involved in the PCP degradation activity of *M. chlorophenolicum* (Das et al. 2015).

In addition to Proteobacteria and Actinobacter, PCP degradation has been identified in members of the phylum Firmicutes. Three strains of the genus *Bacillus* were isolated from contaminated paper mill sludge and were shown to have the ability to completely mineralize PCP (Karn, Chakrabarty, and Sudhakara Reddy 2010).

Pentachlorophenol degradation activity is not confined to bacteria. *Phanerochaete chrysosporium,* a white rot fungus, can completely mineralize PCP as well (Aiken and Logan 1996). Investigations into the mechanism of PCP degradation in this organism implicate the use of cytochrome p450 in the initial hydroxylation reaction (Ning and Wang 2012).

PCP degradation is clearly widespread across multiple domains of life and across many phyla within those domains. Despite the ubiquity of organisms with PCP degradation activity, research identifying the details of the degradation pathways is rare. Among PCP degraders, the degradation pathway of *S. chlorophenolicum* is by far the most extensively characterized.

Uncovering the PCP Degradation Pathway of *S. Chlorophenolicum*

Characterizing the PCP degradation pathway of *S. chlorophenolicum* required the concerted efforts of many researchers for over 30 years. The first enzyme in this pathway was found by identifying a periplasmic protein that was enriched in crude periplasmic protein extract of cells after treatment with PCP. This protein, designated PcpA*,* was purified and N-terminally sequenced. The amino acid sequence was used to design degenerate primers and clone the gene encoding the enzymes from the bacterial genome. Enzymatic assays showed that PcpA is an extradiol dioxygenase that cleaves the aromatic ring of a partially dehalogenated intermediate of pentachlorophenol degradation (L. Y. Xun and Orser 1991; Xu et al. 1999; Machonkin et al. 2010).

The second gene identified, *pcpB,* was found by fractionating crude protein extract and following the hydroxylation activity of PCP into tetrachlorohydroquinone (TCHQ). This activity was assigned to a single protein product that was N-terminally sequenced and used to design

degenerate primers to clone the gene encoding it (L. Xun and Orser 1991; Orser et al. 1993). This gene encoded a monooxygenase responsible for catalyzing the first reaction in PCP degradation.

The same strategy was used to determine the next enzyme, *pcpC*. PcpC catalyzes two successive dehalogenation reactions converting TCHQ to trichlorohydroquinone (TriCHQ) and then into dichlorohydroquinone (DCHQ). This activity was followed through increasingly stringent fractionations and assigned to a single unidentified enzyme. (L. Xun, Topp, and Orser 1992; Habash et al. 2002; Kiefer, McCarthy, and Copley 2002; Warner, Lawson, and Copley 2005). The final product of PcpC was found to be the substrate of the first identified enzyme PcpA.

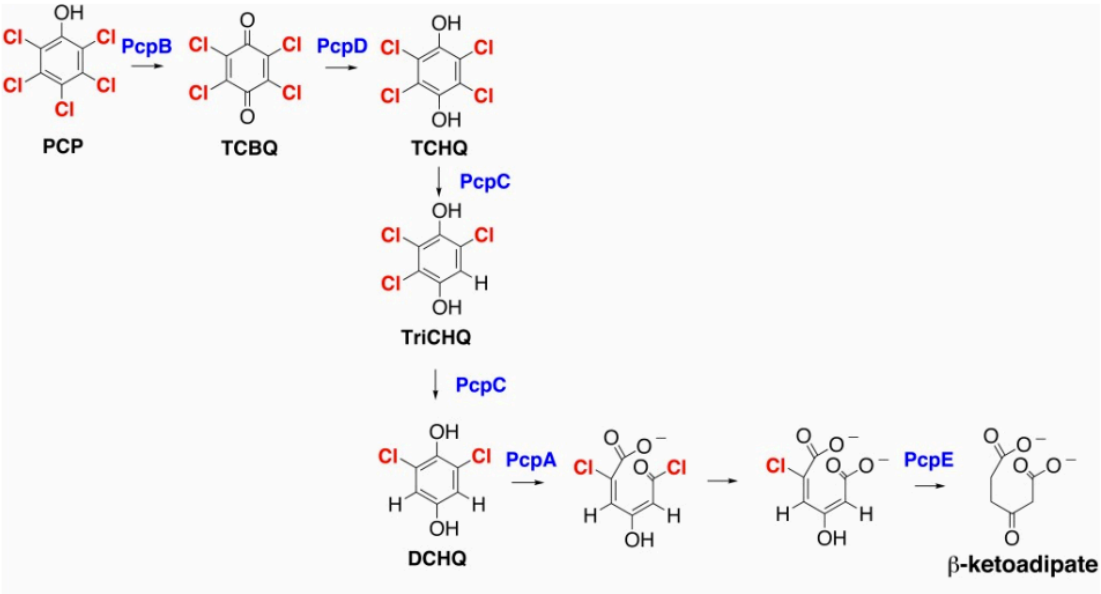
The genomic regions flanking *pcpB* were sequenced and examined for open reading frames. Two genes downstream of *pcpB* were suspected of playing a role in PCP degradation. These genes were named *pcpD* and *pcpR*. PcpD was initially assumed to be the reductase component of a two component oxygenase system including PcpB (McAllister, Lee, and Trevors 1996). Later work, however, established that PcpD directly reduces the product of PcpB and not the PcpB enzyme itself. Hence, the product of PcpB was shown to be tetrachlorobenzoquinone (TCBQ) and not tetrachlorohydroquinone (TCHQ) as previously assumed (Dai et al. 2003). The second gene, *pcpR,* encoded a putative LysR type transcriptional regulator and was presumed to play a regulatory role in inducing the enzymes of the pathway.

The final enzyme in the pathway, PcpE*,* was identified when a 24 kb region including *pcpA* and *pcpC* was sequenced (Cai and Xun 2002). The *pcpE* gene was predicted to encode a maleylacetate reductase. The gene was cloned and the protein was purified and shown to

catalyze the dechlorination of 2-chloromaleylacetate (2-CMA) to maleylacetate and then the reduction of maleylacetate to β-ketoadipate, a common intermediate in the degradation of many aromatic compounds. Metabolic pathways converting β-ketoadipate into intermediates of the TCA cycle were already well characterized.

Thus, over 20 years, the complete enzymatic pathway converting PCP to β-ketoadipate was elucidated. The catabolism of PCP occurs though the successive action of PcpB, PcpD, two successive reactions of PcpC, PcpA, and finally two successive reactions of PcpE, ultimately generating β-ketoadipate that is then funneled into the TCA cycle and used to generate energy for the cell (see Figure 1.2). The elucidation of this pathway begs the question of how such an intricate and specialized pathway could arise in such a short span of time.

Figure 1.2 - The pentachlorophenol degradation pathway



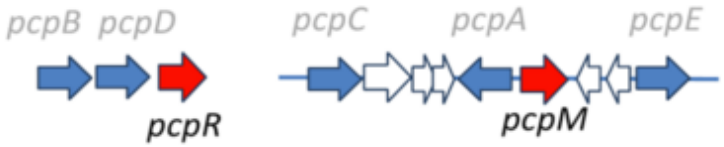
Through the action of five enzymes pentachlorophenol is completely dehalogenated and the aromatic ring cleaved. The resulting product, β-ketoadipate, is then broken down into TCA cycle intermediates.

Genetic Organization and Regulation of the PCP Degradation Pathway

Unlike many degradative pathways in bacteria, the PCP genes are not transcribed in a single operon. In fact, the five catalytic enzymes are transcribed from four separate promoters with only *pcpB* and *pcpD* transcribed in one operon. Three of the four promoters, *pcpBD, pcpA,* and *pcpE,* are strongly induced in the presence of PCP. *pcpC*, however, is expressed constitutively in the absence of PCP but expression increases slightly in the presence of PCP (Rokicki unpublished).

Putative transcription factors were found near the genes encoding the PCP degrading enzymes. The PCP degrading enzymes are located in two clusters in the genome of *S. chlorophenolicum* (see Figure 1.3). Two genes encoding putative regulatory proteins were found to be adjacent to these known clusters of PCP degrading genes. These two regulatory genes were designated *pcpM* and *pcpR*.

Figure 1.3 - Genetic organization of PCP degradation genes



Five PCP degradation enzymes (blue) and two LysR type regulators (red) located in to distinct clusters in the genome of S. chlorophenolicum.

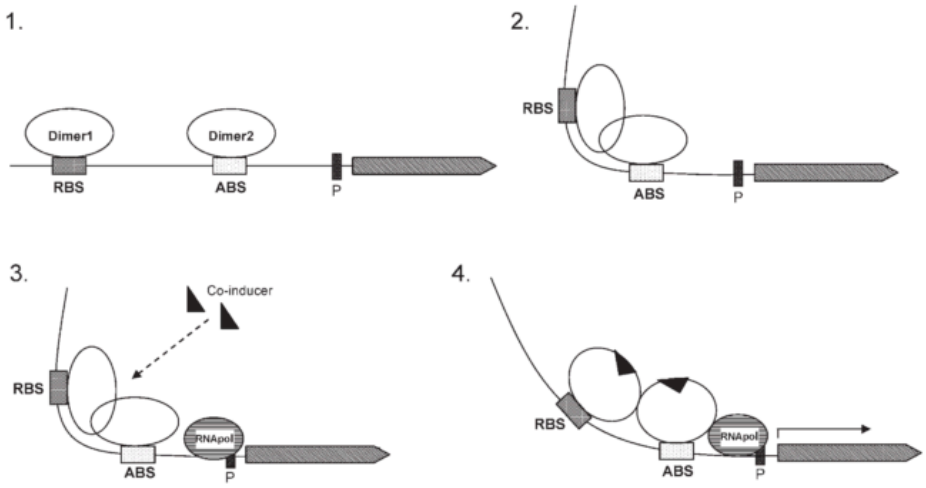
To elucidate the role of PcpR and PcpM in the pathway, knockout strains were constructed and assayed for PCP degradation activity (Cai and Xun 2002). When *pcpR* was disrupted, PCP degradation was completely ablated. Furthermore, induction of the *pcpB* and *pcpE* genes in the presence of PCP was similarly ablated as shown by qualitative RT-PCR. In contrast, when *pcpM* was disrupted, PCP degradation continued to occur at wild type rates.

Induction of *pcpB* and *pcpE* by PCP was similarly unaffected as shown by qualitative RT-PCR. The researchers concluded that *pcpR* was the regulator of the induced PCP genes and that *pcpM* was not critical for PCP degradation. Additionally, the promoter regions of the degradation enzymes were manually scanned for a putative LysR type transcription factor binding site. The LysR type motif ATTC-N_7-GAAT was found upstream of *pcpBD, pcpA* and *pcpE*.

LysR Type Transcriptional Regulators (LTTRs)

The implication of LTTRs in this pathway is consistent with the observation that they are frequently involved in the regulation of aromatic degradation pathways (Tropel and van der Meer 2004). LTTRs function differently from many activators and repressors of transcription. While there are many exceptions, the typical LTTR is constitutively expressed and bound to DNA. The gene of the LTTR itself is often located adjacent and divergently oriented from the gene it is regulating so that the promoters of the two are in the same intergenic region. The LTTR binds in the promoter of its own gene, which is upstream of the promoter of the divergently oriented target gene. LTTRs often function as a dimer of dimers (see Figure 1.4). One dimer binds to a specific NTNN-N_7-NNAN inverted repeat motif called the recruitment binding site (RBS). The second dimer binds a more degenerate version of the same motif labeled the activating binding site (ABS). The bound LTTR tetramer constitutively represses its own expression by occluding its own promoter while simultaneously conditionally activating the expression of the divergently oriented target gene. In the presence of an inducer, the LTTR dimer of dimers will undergo a conformational change leading to the activation of its target gene while maintaining the constitutive repression of its own promoter.

Figure 1.4 - Schematic of a canonical LysR Type Transcriptional Regulator (LTTR)



Adapted from (Maddocks and Oyston 2008). *1) Two LTTR dimers bind to DNA upstream of the regulated ORF. 2) Two LTTRs form a dimer of dimers. 3) Co-inducer leads to conformational change that recruits or activates RNApol. 4) Active RNApol transcribes the target gene. Not shown: The region occluded by the RBS and ABS often includes the promoter for the divergently oriented LTTR gene leading to its constitutive repression.*

The apparent involvement of two LysR regulators in the PCP degradation pathway is problematic. One theory for the evolution of this pathway, discussed in Chapter 2, is that the upstream *pcpB, pcpD* genes arrived via horizontal gene transfer with their own regulator, the *pcpR* protein, and that the downstream *pcpC, pcpA,* and *pcpE* genes were already present and regulated by *pcpM.* If this were the case, then the two LTTRs would have collided, competing over regulation of the target genes. The situation becomes much more complicated when the regulatory mechanisms of LTTRs are taken into account, specifically, the fact that they are constitutively expressed and tend to negatively auto-regulate themselves. Resolving the regulatory mechanisms associated with PCP degradation and identifying other genes that are similarly regulated would inform the evolutionary history of this pathway.

Conclusion

The minimal pathway necessary to catalyze the degradation of PCP has been discovered but there are many significant gaps in our understanding of both the evolutionary and regulatory relationships these genes share with the rest of the genome. Where did these genes come from? Did they originate by horizontal gene transfer, recruitment, or duplication and divergence? Are they unique in the genome or are they accompanied by close paralogs? Many of the genes are induced by PCP but what other genes in the genome are similarly induced? Through genome sequencing, transcriptomics, and biochemical studies of the transcription factors involved, I can uncover the evolutionary and regulatory relationships between the PCP degradation pathway enzymes and the rest of the genome. These relationships create a context that informs the evolutionary history of this unique pathway

Chapter 2 Genome Sequencing of *S. chlorophenolicum*

Part of this chapter was published as (Copley et al. 2012). My contribution to this publication were computational analysis and figure generation for figure 2, figure 3, figure 4, figure 5, and figure 6 (numbering as in Copley et. al. 2012) and both supplementary tables. I performed all computational and statistical analysis described in the paper and conceived of the idea of a comparative genomic analysis between *S. chlorophenolicum* and *S. japonicum.* The paper was written by Shelley Copley.

Introduction

One of the leading theories for the generation of genes with new functions is duplication and divergence (Bergthorsson, Andersson, and Roth 2007). A major motivation for sequencing the *S. chlorophenolicum* genome was to uncover paralogs of the PCP degradation enzymes. Additionally, genome sequencing would facilitate many types of experiments that were previously impossible ranging from the sophisticated such as RNAseq to the more mundane such as designing primers. In this chapter, I describe the process of genome sequencing and annotation and I perform many types of primary sequence analysis and comparative genomics.

Genome Sequencing of *S. Chlorophenolicum*

We sequenced the genome of *S. chlorophenolicum* L-1. Analysis of this sequence and comparison with the sequence of the closely related *Sphingobium japonicum*, which degrades lindane (Nagata et al. 2010), provides insights into the origins of the PCP degradation enzymes.

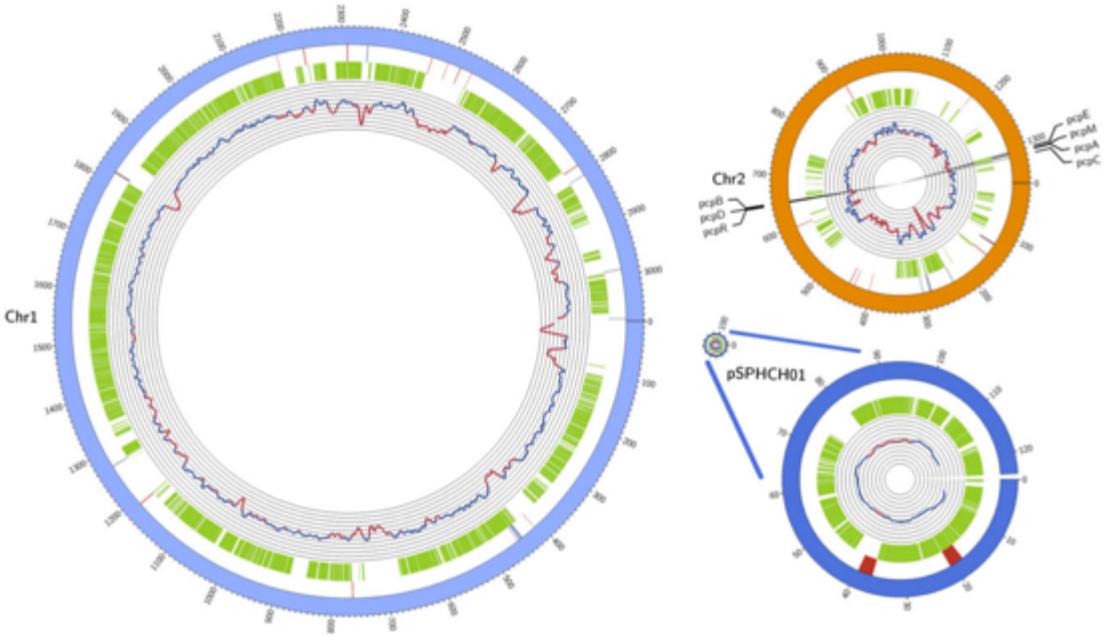
The rDNA genes of *S. chlorophenolicum* and *S. japonicum* share 97% identity. The phylogenetic distance between *S. chlorophenolicum* and *S. japonicum* is ideal to facilitate identification of genes that were present in the most recent common ancestor of these two Sphingomonads, as proteins involved in core processes show >80% pairwise identities at the amino acid level.

Our analysis suggests that the first three enzymes in the pathway were acquired by *S. chlorophenolicum* by horizontal gene transfer (HGT) after it diverged from *S. japonicum*. In contrast, the last two enzymes in the pathway were present in the most recent common ancestor of *S. chlorophenolicum* and *S. japonicum.* None of the genes encoding the PCP degradation enzymes arose by recent duplication and divergence of genes within *S. chlorophenolicum*. The genes occur in two disparate parts of the genome and have not yet been integrated into a compact and consistently regulated operon.

Overview of the Genome

The *S. chlorophenolicum* genome consists of two chromosomes and a plasmid (see *Figure 2.1*)

Figure 2.1 - The *S. chlorophenolicum* replicons.

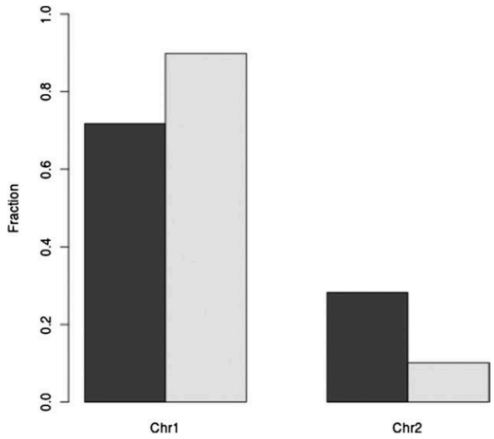


Three circular diagrams represent the two chromosomes and one plasmid that make up the Circle diagram of the replicons of S. chlorophenolicum. The second circle in each replicon indicates the locations of PCP degradation genes in black, phage genes in blue, and transposon-related genes in red. The third circle shows (in green) locations of genes that have a close homolog in S. japonicum (80% identity over 90% of the S. chlorophenolicum sequence). The fourth circle shows GC content. Red indicates sequences with GC content <64% and blue indicates sequences with GC content >64%. Light gray lines are placed at intervals of one standard deviation from the mean of 63.8%. (One standard deviation is 0.027%; GC content was calculated by averaging over a 10-kb window and sliding that window in 1-kb increments.) Diagram was generated using Circos.

Like many bacteria that contain multiple replicons such as *Burkholderia pseudomallei* (Holden et al. 2004) and *Vibrio cholera* (Heidelberg et al. 2000), *S. chlorophenolicum* has a dominant chromosome which is contains most of the essential genes. We generated a list of genes involved in replication, transcription, translation, cell division, peptidoglycan biosynthesis, and core metabolic processes (including glycolysis, the pentose phosphate pathway, the TCA cycle, electron transport, and biosynthesis of amino acids, nucleotides, and cofactors). Figure 2.2 shows the distribution of genes between the two chromosomes. If genes encoding core

functions were randomly distributed between the chromosomes, we would expect 354 (72%) to be on chromosome 1 and 138 (28%) to be on chromosome 2. In fact, 442 core genes (90%) are on chromosome 1 and only 50 (10%) are on chromosome 2. This deviation from the expected distribution is highly significant by chi-square analysis (p-value = 2.2 * 10^-16).

Figure 2.2 - Fraction of core metabolism genes on Chr1 and Chr2

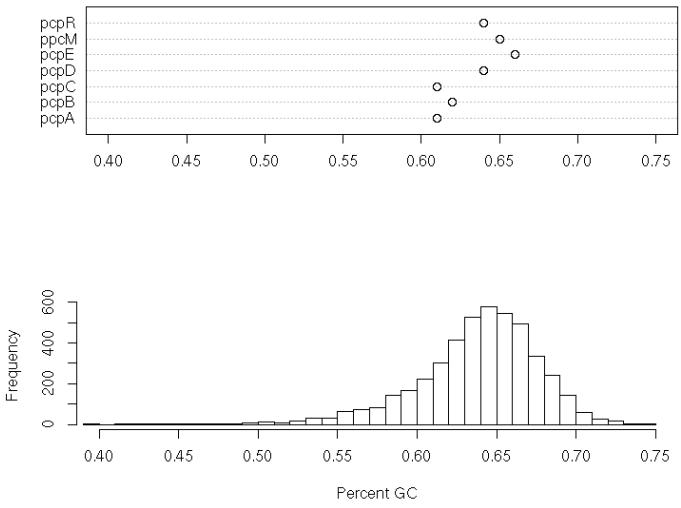


The 50 genes on chromosome 2 include an odd collection of essential enzymes, including a few genes for central carbon metabolism, two subunits of RNA polymerase, some components of the electron transfer chain, and a few genes for cofactor biosynthesis. However, most of the genes on chromosome 2 appear to be involved in environmental adaptation. Chromosome 2 also carries a number of genes predicted to encode transporters for sugars and metal ions and enzymes involved in degradation of various sugars, short-chain fatty acids and aromatic compounds. All of the PCP degradation genes are found on chromosome 2. A secondary chromosome with a low density of essential genes may serve as a convenient storage depot for genes acquired by HGT, as integration of newly acquired DNA is not likely to disrupt an essential function. Chromosome 2 also contains two genes encoding proteins related to ParB, which is involved in plasmid partitioning. These features of chromosome 2 are

consistent with the proposal that bacterial secondary chromosomes have arisen by intragenomic transfer of essential genes, including rRNA genes, to a plasmid (Slater et al. 2009).

All of the PCP genes are encoded on chromosome 2. I looked to see if they had anomalous GC content and found that they do not (see Figure 2.3).

Figure 2.3 - GC Content of the PCP Genes

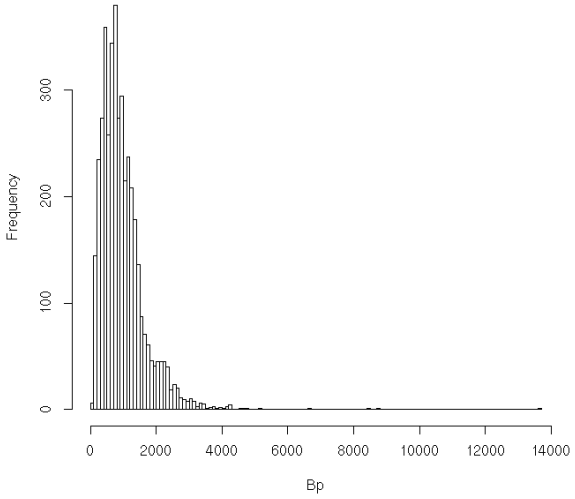


The GC content of the PCP genes (top panel). Histogram of the GC content of all genes in S. chlorophenolicum.

Open Reading Frame Analysis

ORFs were annotated as described in the methods. The size distribution of all ORFs in *S. chlorophenolicum* is plotted in the Figure 2.4.

Figure 2.4 - The size distribution of ORFs in S. chlorophenolicum



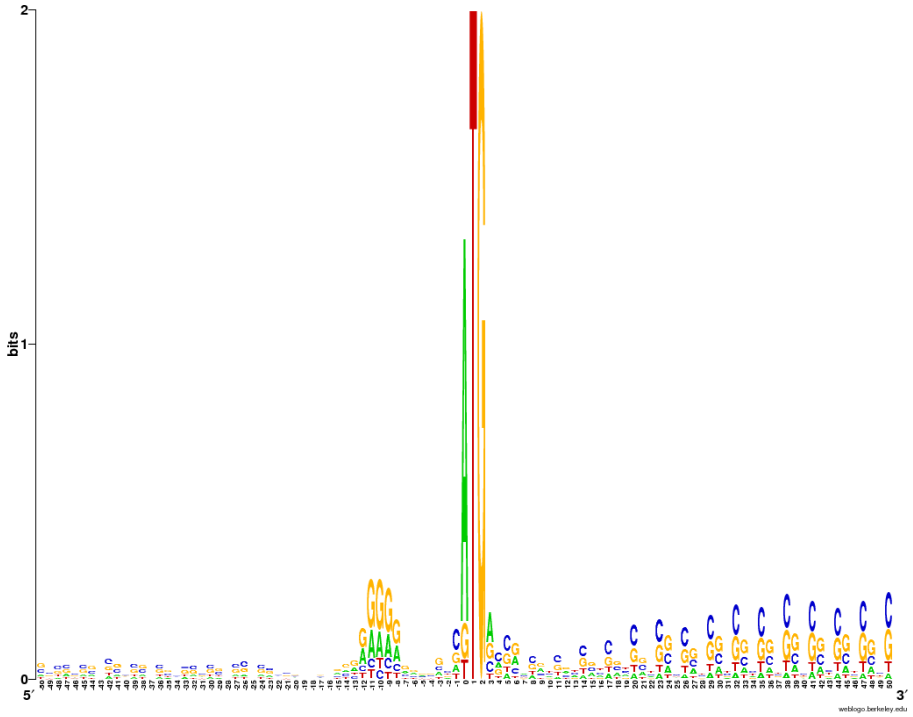
Histogram of the bp length of all genes encoded by the S. chlorophenolicum genome

The average ORF has a size of 976 bp. There is a large standard deviation and the distribution is non-gaussian. There is a large tail of larger than average ORFs with the largest annotated ORF being more that 13kb. This ORF size distribution is similar to that of other bacteria.

Ribosome Binding Sites

After precisely identifying the location of the ORFs in *S. chlorophenolicum* I investigated if this organism utilized non-standard ribosome binding sites and if there were any other unexpected sequence biases in the vicinity of the start codon of an average gene. To test this, I pulled 100 bp sequences centered on the start codon of each ORF such that position 0 was the A in the ATG of the start codon. These sequences were then converted into a sequence logo below using WebLogo (Crooks et al. 2004).

Figure 2.5 - RBS and codon bias of the average S. chlorophenolicum ORF.



Consensus sequence of a 100bp window centered on the start codon of every ORF encoded by the S. chlorophenolicum genome

The ATG is present in almost 100% of the sequences aligned as represented by their 2 bits of information. The gene annotation algorithm, Prodigal, allows for alterative start codons such as GTG and TTG in some cases. A small hump of GA-rich sequence is apparent at -10 base pairs upstream of the start codon. This corresponds to the ribosome binding sequence. There is also a periodic GC bias that appears downstream of the start codon and repeats every third base pair. This bias is present at the "wobble" position or third position of every codon downstream of the start codon. A secondary but smaller periodic GC enrichment is apparent in the first position of each codon. Both of these biases increase with increasing distance from the start codon.

This periodic GC bias can be explained as a manifestation of the codon bias of the organism. *S. chlorophenolicum* is a GC rich organism and much of that GC richness is explained

by a preference for choosing a G or C in the wobble position of most codons. It is interesting that this signal from codon bias is diminished in the four or five codons adjacent to the start codon. I interpret this as a selection pressure preventing GC rich sequences this close to the ribosome binding site that could cause mRNA secondary structures that inhibit translation (Kosuri et al. 2013). Finally, there is a faint periodic GC bias upstream of the ribosome binding sequence as well. This could be the codon bias signal of upstream ORFs in polycistronic mRNAs.

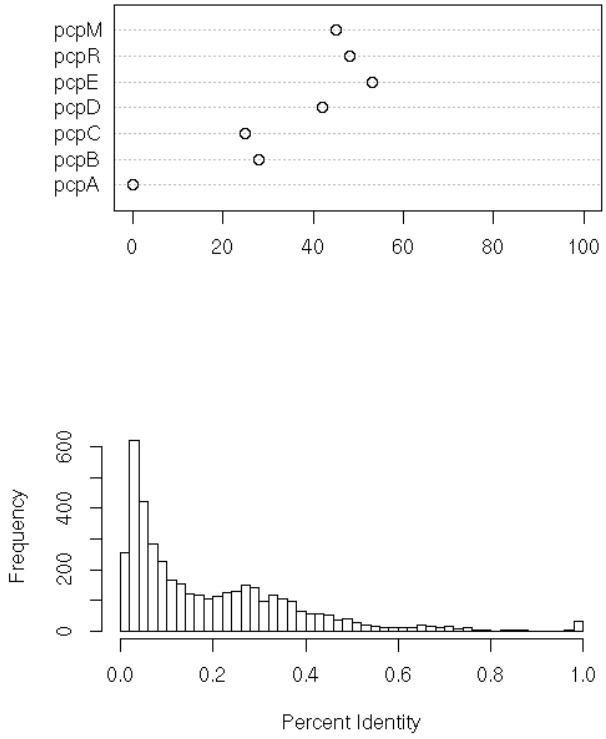
Paralogs

New genes that can help microbes survive and grow in the face of selective pressure from environmental toxins can also arise by gene duplication and divergence (Bergthorsson, Andersson, and Roth 2007; Hughes 1994). I carried out a BLAST search of the *S. chlorophenolicum* genome against itself to identify genes that are nearly identical and may have arisen by recent gene duplication. I found that only 38 proteins have >90% sequence identity to another protein in *S. chlorophenolicum*. Of these, 13 are related to transposases; three of these genes are found in three identical copies scattered throughout the genome. While the presence of highly similar transposase genes is not unexpected, there is little rhyme or reason to the identities and locations of the remaining duplicated genes. Seven genes annotated as encoding 2-hydroxychromene-2-carboxylate isomerase, a short-chain alcohol dehydrogenase, a hypothetical protein, a nucleoside-diphosphate-sugar epimerase, an arabinose efflux permease, a Zn-dependent dipeptidase and an outer membrane receptor protein, are present between positions 667462 and 675176 on the – strand of chromosome 2. A second copy of five of these genes are found in a cluster on the + strand of chromosome 2 between positions 1036969 and 1030812. Copies of the first two genes in the cluster are found together in a different location

on chromosome 2. Nearly identical copies of the genes encoding the alpha and beta subunits of the E1 component of pyruvate dehydrogenase are found on chromosomes 1 and 2, and nearly identical copies of a gene annotated as an acyl-CoA synthetase/AMP-acid ligase II are found in chromosome 2, but surrounded by completely different genes. The utility of these duplicated genes and the processes leading to their location in distant parts of the genome are not clear. One possibility is that they were not actually duplicated in *S. chlorophenolicum*, but rather acquired by integration of DNA fragments taken up from lysed cells of *S. chlorophenolicum* or closely related Sphingomonads. Notably, none of the PCP degradation genes are found among the set of highly similar genes.

It is worth noting that although I did not find homologs close enough to be designated paralogs, I did find two more distantly related homologs of *pcpE* in the *S. chlorophenolicum* genome. Each homolog shares about 50% identity. In *S. japonicum,* these *pcpE* homologs are absent but there are two homologs of *pcpA*. This suggests either an ancient duplication event in the common ancestor of these strains or two independent and equally ancient duplication events. The *pcpA* ortholog in *S. japonicum* retains the 2,6-dichlorohydroquinone substrate activity of the ortholog in *S. chlorophenolicum*. The other homolog has 2,5-dichlorohydroquinone activity associated with its role in the degradation of lindane, as well as limited 2,6-dichlorohydroquinone activity (Endo et al. 2007).

Figure 2.6 - Histogram of paralogous relationships in S. chlorophenolicum



Histogram of the percent alignment of each genes best hit within the genome of S. chlorophenolicum. Percent identity on the x-axis. Counts on the y-axis.

The histogram of paralogous percent identities appears to have four distinct peaks. The largest peak at around ~10% likely corresponds to spurious low p-value relationships between genes without true paralogs in the genome. The second peak at ~30% likely corresponds to homology relationships where proteins of the same family share sequence identities but are not true paralogs recently duplicated in the *S. chlorophenolicum* genome. Finally, there are two more peaks at around 65% and at 100%. These are likely true paralogs that have duplicated recently or more anciently and have been maintained in the genome. The 50% sequence identity of the three *pcpE* homologs puts them right at the border of homologous relationships and more ancient duplications.

Codon usage and GC content

Different organisms favor the usage of different synonymous codons. I examined the *S. chlorophenolicum* codon bias and found that the majority of it can be explained by a GC bias in the wobble position. Interestingly, this same phenomenon explains the high GC content of the organism. When the GC content measurement is confined to intergenic regions, it drops markedly.

Mobile Elements: Prophage and Transposon Insertions

HGT is rampant among microbes and is known to play a major role in acquisition of resistance to or degradation of toxic compounds, including antibiotics. I examined the *S. chlorophenolicum* genome for features indicative of mobile genetic elements. As mentioned above, *S. chlorophenolicum* contains one plasmid. The genome appears to contain one integrated pro-phage; a cluster of several phage-related genes (including a major capsid protein, major tail protein, phage portal protein, pro-head peptidase and some conserved phage proteins of unknown function) is found on chromosome 1 (genes Sc_00004260-00004380). Curiously, seven isolated genes annotated as "phage integrase family" genes are found on both chromosomes 1 and 2 (Chr1: Sc_00026840, Sc_00029370, Sc_00022480, Sc_00028520, Sc_00012030; Chr 2: Sc_00031410, Sc_00030440. These genes have anomalously low GC content (0.48 – 0.58) and are not closely related to each other. Only one (Sc_00026840) is found in the vicinity of a prophage gene (a CP4-57 regulatory protein). Two (Sc_00017260 and Sc_0003440) are adjacent to transposase genes. The genome also carries 27 sequences annotated as "transposase", "transposase/integrase core domain", or "transposase and

inactivated derivatives" (16 on chromosome 1, 9 on chromosome 2, and 2 on the plasmid). These transposons belong to several families, including the IS3/IS911, IS30-like.

Thus, like most microbial genomes, the genome of *S. chlorophenolicum* displays evidence of continual onslaught by mobile genetic elements. However, the PCP degradation genes show no association with any of these elements.

Core metabolism genes are statistically enriched on chromosome 1

Just as horizontally transferred genetic material appears to be preferentially present on chromosome 2, we noticed that genes involved in core metabolism are preferentially located on chromosome 1. We annotated a list of genes involved in core metabolic and cellular processes. Genes from this list appear on both chromosomes, but they are very significantly overrepresented on chromosome 1 (see Figure 2.2).

There are several ways to interpret this correlation. It could be a result of the previous observation that chromosome 2 is a landing pad for horizontal gene transfer. If there is some mechanism for preferentially incorporating foreign DNA into chromosome 2, then having the essential genes on chromosome 1 provides a selective advantage, as horizontal gene transfer is less likely to disrupt a core process. Alternatively, this bias could just as likely be the cause as the result of the chromosomal bias of HGT. Perhaps, the bias of core genes to chromosome 1 makes HGT in chromosome 1 more likely to result in disrupting an essential process and therefore leading to a bias of HGT to chromosome 2.

Comparative Analysis Of The *S. Chlorophenolicum* Genome

The genomes of *S. chlorophenolicum* and *S. japonicum* are of similar size (4.57 and 4.46 Mbp, respectively), and contain a similar number of ORFs (4159 and 4460, respectively). Both *S. chlorophenolicum* and *S. japonicum* (Nagata et al. 2010) have a primary chromosome containing most of the genes for core processes (including glycolysis, the TCA cycle, amino acid and nucleotide biosynthesis, fatty acid oxidation, DNA replication, transcription and translation) and a secondary chromosome. *S. chlorophenolicum* has a single plasmid (pSphCh01), while *S. japonicum* has three (pUT1, pUT2 and pCHQ1).

The third circle in each chromosome map in *Figure 2.1* shows in green the positions of genes that encode proteins in *S. chlorophenolicum* that have close homologs in *S. japonicum* that exhibit >80% identity over >90% of the length of the *S. chlorophenolicum* sequence. A total of 2324 *S. chlorophenolicum* genes (1931 on chromosome 1, 285 on chromosome 2 and 108 on pSphCh01) have close homologs in *S. japonicum* (see Supplementary Table 2). In both chromosomes, regions with close homologs in *S. japonicum* show a typical GC content of about 64%. Regions without close homologs (*S. chlorophenolicum* islands) might have resulted from either loss of genes in *S. japonicum* or acquisition of genes by HGT in *S. chlorophenolicum*. Some of the *S. chlorophenolicum* islands show a lower GC content, suggesting that these regions may have been acquired by HGT. Most of the genes in the *S. chlorophenolicum* islands are hypothetical proteins, but there are several predicted glycosyltransferases (some predicted to be involved in cell wall biosynthesis), as well as some O-antigen ligases, some ABC transporters, cellobiose phosphorylase, a K+-transporting ATPase, and Type IV secretory

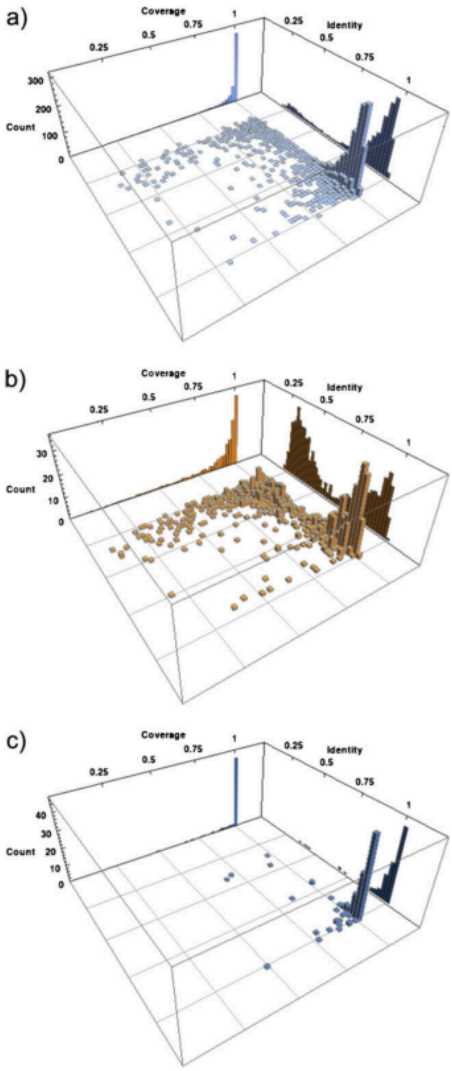
pathway components. Notably, almost all of the sequences associated with transposons and phage genes are found in *S. chlorophenolicum* islands.

A more detailed analysis of the relationships between proteins found in both *S. chlorophenolicum* and *S. japonicum* is shown in Figure 2.7, which shows plots of sequence identity vs coverage for the top hit in the *S. japonicum* genome for each *S. chlorophenolicum* protein. Coverage is defined as the length of the *S. chlorophenolicum* query sequence that is aligned to a sequence in *S. japonicum* divided by the total length of the query sequence. The data were filtered to remove pairs for which the e-value was > 0.0001. On this plot, homologs cluster in three regions: 1) close homologs that share high sequence identity over most of the query sequence; 2) more distant homologs that share moderate sequence identity over most of the query sequence; and 3) homologs that share sequence identity only over part of the query sequence. Notably, chromosome 1 is highly enriched in close homologs, and chromosome 2 is modestly enriched in distant homologs. Considered with the observation that most of the genes for core metabolic processes are present on chromosome 1, this observation suggests that chromosome 2 may preferentially collect horizontally transferred genes.

Most of the proteins encoded on chromosome 1 share > 80% identity with proteins in *S. japonicum* (see Figure 2.7); indeed, 65% share > 90% identity. I posit that close homologs with >80% identity were present in the most recent common ancestor of these two species; most are likely to be orthologs. The more distant homologs in region 2 are unlikely to be orthologs derived from the most recent common ancestor, since they are much more divergent than the large number of close homologs in region 1. The genes encoding these proteins may have been acquired by HGT independently from different sources in the two species; they may serve the

same or different functions. Alternatively, the *S. chlorophenolicum* protein may indeed have derived from the most recent common ancestor, but the ortholog in *S. japonicum* may have been lost so that the best hit in the *S. japonicum* genome is actually a paralog. Finally, there are 112 proteins for which significant sequence identity is seen over only part of the query sequence (<75%); these include proteins in which conserved domains have been utilized in different structural contexts.

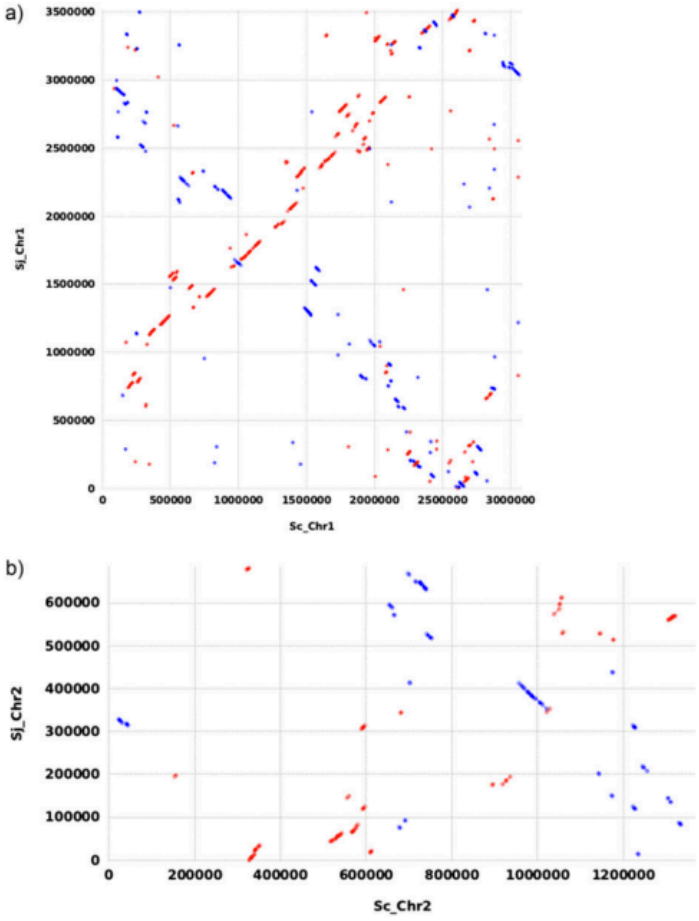
Figure 2.7 - S. chlorophenolicum vs S. japonicum homologous proteins.



Histogram of coverage and identity of orthologs between S. japonicum and proteins from a) Chr1 b) Chr2 or c) Plasmid 1 of S. chlorophenolicum.

Although the dominant chromosomes of *S. chlorophenolicum* and *S. japonicum* share a common core of genes, there has been considerable rearrangement of genes since the common ancestor of these two bacteria. Figure 2.8 shows a scatter plot of gene conservation between the chromosomes of *S. chlorophenolicum* and *S. japonicum* made using nucmer in the MUMmer package with the default parameters (Delcher, et al. 1999; Delcher, et al. 2002; Kurtz, et al. 2004).

Figure 2.8 - X-alignment conservation between S. chlorophenolicum and S. japonicum

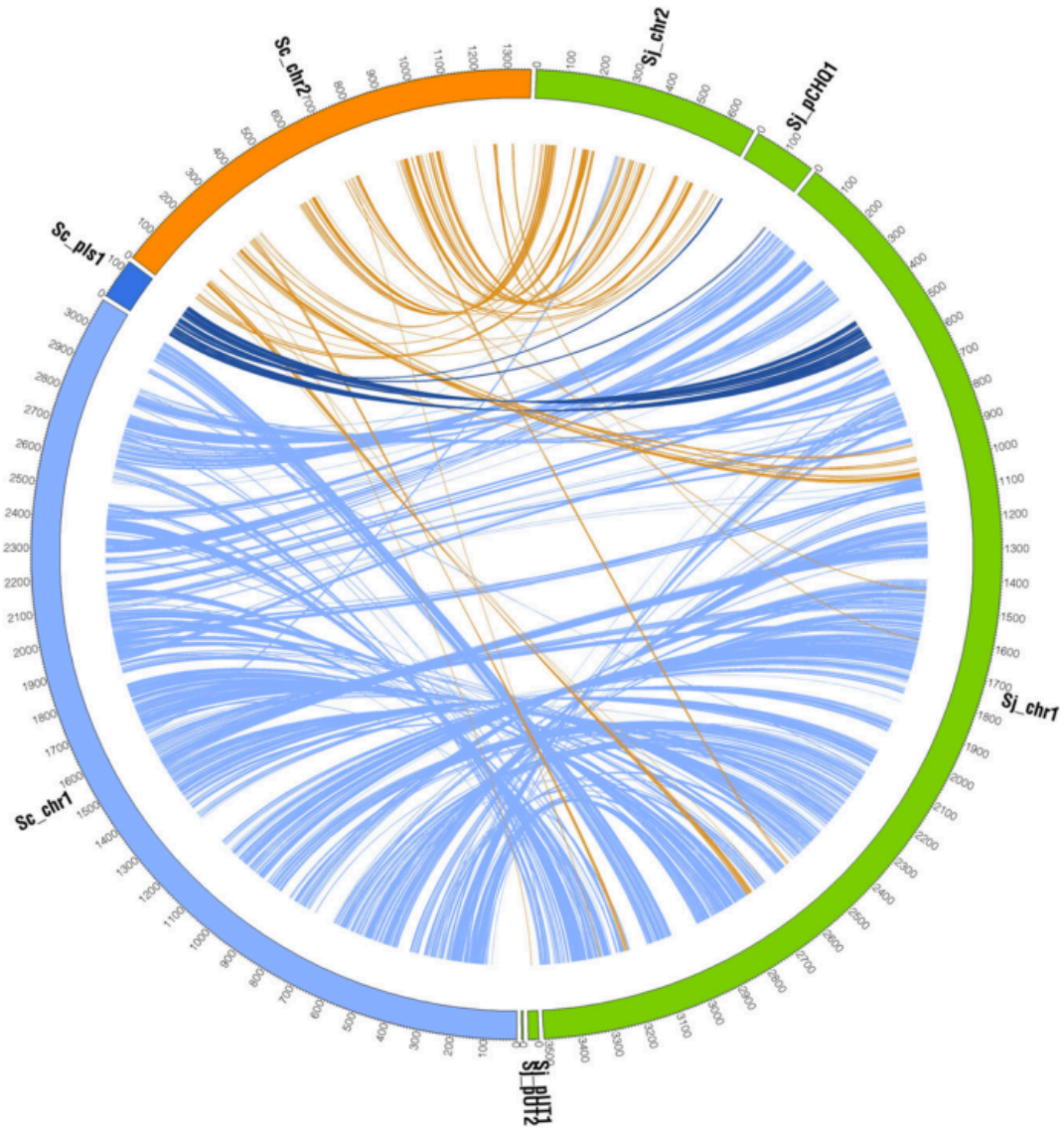


Dotplot representing regions of homology between the orthologous chromosomes of S. chlorophenolicum and S. japonicum. The horizontal and vertical axes are the bp positions of the aligned chromosomes. Direct alignments are plotted in red. Inverted alignments are plotted in red. Panel A is an alignment of chromosome 1 of each species. Panel B is an alignment of chromosome 2 of each species.

This plot shows a pattern known as an “x-alignment” (Eisen, et al. 2000) that is often seen in alignments between closely related bacteria. Many genes in chromosome 1 of *S. chlorophenolicum* are found in comparable positions in chromosome 1 of *S. japonicum*, as indicated by the red dots along the diagonal. However, a number are found in an inverted orientation (see blue dots) in positions that lie close to a diagonal perpendicular to the red diagonal. This pattern is believed to result from multiple inversions centered on either the origin or terminus of replication. There has evidently been considerable remodeling of the genome via movement of blocks of genes within chromosome 1 since *S. chlorophenolicum* and *S. japonicum* diverged from a common ancestor. The x-alignment pattern is less distinct for chromosome 2, suggesting greater plasticity in this replicon. Again, this is consistent with the lower number of genes for core processes on chromosome 2.

Additionally, Figure 2.9 depicts the correspondence between the positions of very close homologs (>90% identity over >80% of the query length) in the entire genomes of *S. chlorophenolicum* and *S. japonicum*.

Figure 2.9 - Orthologous proteins in S. chlorophenolicum and S. japonicum



Each colored segment in the outer ring represents of replicon of either S. chlorophenolicum of S. japonicum. Ribbons connect the orthologs of the two species with >90% identity and >80% coverage. The ribbons are colored by the S. chlorophenolicum ortholog replicon.

Light blue, orange and dark blue lines connect the positions of genes in chromosome 1, chromosome 2 and the plasmid, respectively, of *S. chlorophenolicum* with the positions of homologs in the two chromosomes and three plasmids of *S. japonicum.* As noted above, chromosome 1 in both species is densely populated with shared genes, shown in light blue. A

number of genes found on chromosome 2 of *S. chlorophenolicum* have homologs on chromosome 1 of *S. japonicum*, although the converse is not true. Notably, genes found in *S. chlorophenolicum* but not in *S. japonicum* are more heavily represented on chromosome 2 than on chromosome 1. Since chromosome 2 appears to carry many of the genes for degradation of organic compounds, this difference may be due to the availability of different carbon sources in the environmental niches occupied by the two bacteria. Notably, the *S. chlorophenolicum* plasmid, pSphCh01, is comprised of a large region that is homologous and syntenic with a region of *S. japonicum* chromosome 1 (with the exception of a few small indels) and a smaller region that is homologous and syntenic with a region around the origin of *S. japonicum* pCHQ1. This region encodes several proteins, including two chromosome partitioning proteins (ParA and ParB homologs) and the plasmid replication initiation protein (RepA). Thus, the *S. chlorophenolicum* pSphCh01 and *S. japonicum* pCHQ1 share an origin of replication and the associated genes, but the genes carried on the two plasmids are not closely related. The two smallest plasmids in *S. japonicum* (pUT 1 and pUT2) carry genes with no homologs in *S. chlorophenolicum*.

Most of the genes on the *S. chlorophenolicum* plasmid are also found on plasmids in *Sphingomonas wittichii* (a more distantly related Sphingomonad that is the closest relative of *S. chlorophenolicum* and *S. japonicum* for which a whole genome sequence is available) and *Sphingobium* SYK-6 suggesting that these genes may have been present on a plasmid in the ancestor of *S. chlorophenolicum* and *S. japonicum*, and may have been incorporated into chromosome 1 of *S. japonicum* after divergence of the two species. Movement of blocks of genes among the plasmids and chromosomes in these organisms may have been facilitated by

transposases, as TN3-family transposase elements are present in both plasmids and the *S. japonicum* chromosome near one end of the integrated region.

Conclusion

Through genome sequencing and comparative genomics I elucidated the relationship of the PCP degrading enzymes to the rest of the genome. I was surprised to find an absence of close paralogs for any of the PCP genes likely ruling out a history of duplication and divergence in favor of recruitment. *pcpE* is accompanied by two distant homologs of about 50% sequence identity. The ortholog of *pcpA* in the genome of *S. japonicum* also has a homolog that shares about 50% sequence identity in that genome. At this evolutionary distance, it is impossible to tell if these homologs duplicated in the ancestral genome or diverged in separate lineages and combined by HGT. However, the evolutionary pattern is consistent with a process of ancient duplication and divergence in the ancestor of these two species followed by loss of different paralogs in the different lineages.

I found that core metabolism genes are preferentially located on the primary chromosome and that the secondary chromosome contained more genes for putative degradation pathways and is thus hypothesized to be more frequently visited by horizontal gene transfer.

Finally, I found evidence of anomalous %GC content in the *pcpBD*, *pcpR* cluster of genes suggestive of an HGT event, but I did not see any repetitive or transposase elements nearby to suggest a mechanism for the horizontal transfer.

Chapter 3 CodaChrome tool for proteome comparisons.

Part of this chapter was published as (Rokicki et al. 2014). My contribution to this publication was designing and implementing the program, writing and publishing its use documentation online, utilizing it to find the examples listed in the paper, and writing the paper.

Introduction

After sequencing the genome of *S. chlorophenolicum,* I faced the task of the analysis described in the previous chapter. Many bioinformatics tools and packages were available for primary sequence analysis such as identifying GC content, A/T skew, transposable elements, terminators, etc. I quickly realized that some of the most insight however could be gained through comparative genomics rather than primary sequence analysis. A comparison of *S. chlorophenolicum* to a closely related fully sequenced ancestor, *S. japonicum*, became the foundation for the genome paper we wrote.

In contrast to tools for primary sequence analysis, there were relatively few tools for comparative genomic analysis. Likely, because until relatively recently there were only a handful of fully sequenced bacterial genomes. As the number of fully sequenced bacterial genomes increases exponentially this type of analysis becomes increasingly powerful. The need for fast, user friendly, comparative genomics tools for the analysis of bacteria is obvious.

The relationships between bacterial genomes are complicated by rampant horizontal gene transfer, varied selection pressures, acquisition of new genes, loss of genes, and divergence of genes, even in closely related lineages. As more and more bacterial genomes are

sequenced, organizing and interpreting the incredible amount of relational information that connects them becomes increasingly difficult

Mauve

One tool, Mauve (Darling et al. 2004), was used frequently when trying to get a handle on how the genome of *S. chlorophenolicum* related to the genome of other microbes. This program would perform a whole genome alignment between two or more genomes and generate an interactive plot with segments of contiguous orthologous genome boxed and colored. A line segment connects the orthologous blocks. This program has support for GenBank files and so gene annotations can be viewed and very quickly the user can ascertain the level of conservation for different genes, operons or whole sections of genomes.

This program works very well for closely related genomes but in bacteria, gene order decays very quickly even between very closely related species. Genes are shuffled like cards in a deck and so even though at the level of protein sequence identity many genes are still very conserved, the synteny is completely rearranged and the result of attempting to draw a line between every orthologous block is a very complicated diagram of limited usefulness.

The second limitation of Mauve is that by the detailed nature of the plots, Mauve graphs are intelligible for comparing only 2 or 3 genomes. Similarly the computational requirements of the multigenome alignment increase exponentially as more genomes are added making the alignment of very many genomes computationally expensive. These two limitations motivated the creation of a new program for high throughput genome comparison I named CodaChrome.

CodaChrome Design Specifications

At the outset of programming CodaChrome, we made very specific prescriptions for what it would and would not be.

The two major limitations of other comparative genome software that I hoped to overcome with CodaChrome were universality and evolutionary depth. I wanted a single visualization that could compare not just one or two genomes but every fully sequenced genome in the database queried. If any protein in the genome of interest has any significant level of conservation with any other protein in any other genome we want that information visualized. Furthermore, I wanted a comparison that conveyed the information of how thousands of fully sequenced genomes relate to a single genome of interest.

I ended up creating a graph that I thought met these goals: the CodaChrome Graph. A CodaChrome graph is a heat map where each colored square along the x axis represents an ORF in the order it appears in the genome of the "seed organism". This seed organism is the lens through which all other genomes will be viewed. The Y axis corresponds to every other fully sequenced genome in GenBank that contains at least a single open reading frame with some significant level of protein sequence identity. The color of a square in the heat map signifies the percent identity shared between the seed organism ORF at that x axis position and the best hit in the other genome specified by that row. A graph like this overcomes the two major limitations described earlier. First of all, the relationship of many thousands of genes to many thousands of genomes can be visualized simultaneously in a single image. By allowing for zooming in and zooming out and rendering the image with a special scaling algorithm, I am able to concisely summarize millions of relationships in a single image. The second limitation of the

rapid syntenic decay in bacteria is also overcome because this method of whole genome comparison is independent of the gene order of any species except for that of the seed organism.

Very quickly many patterns become visible. This will be discussed in the next section. Additionally, the decision to plot protein sequence identities as opposed to nucleic acid identities has two consequences. First, the percent identity of protein alignments remains significant over a much larger evolutionary distance than the percent of nucleic acid alignments. Alignments that would not be significant for DNA are very significant for protein alignments. The bioinformatics signal of protein sequence conservation reaches across greater evolutionary distances than that of DNA sequence conservation.

Finally, I wanted to make the user experience of CodaChrome intuitive for biologists across the entire spectrum of computer literacy. For this reason, I conservatively opted against any menus and restricted the functionality of CodaChrome to a handful of buttons that are always visible on the screen. In several cases, I implemented the same functionality redundantly. For example, a user can zoom in on a particular region by clicking the “zoom” button or by clicking and dragging a rectangle around the region of interest in the overview panel. Dual implementations like this allow users to interact with the program in the way in which they are most comfortable.

Similarly, I wanted CodaChrome to function on whatever operating system the biologist is most comfortable with. I decided to write and compile the program so that it could be run on the latest versions of the three current major operating systems: Linux, Mac OS, and Windows.

Finally, I wanted the program to load and run quickly even on outdated hardware regardless of graphics card capabilities.

Implementation Of CodaChrome

Generation of the CodaChrome matrix file

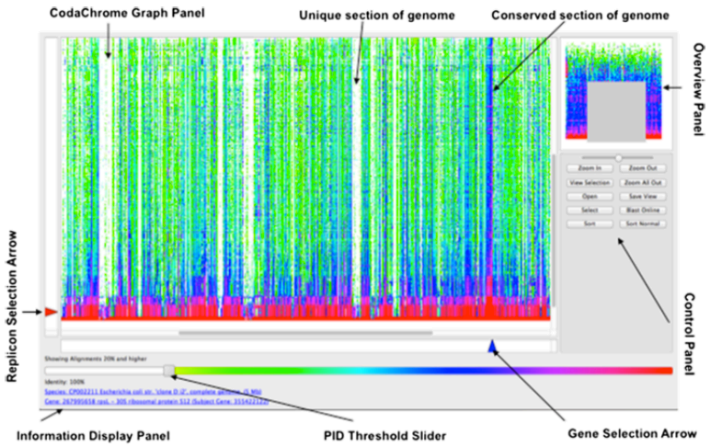
CodaChrome consists of a series of PERL scripts that generate a CodaChrome matrix file and a user-friendly graphical user interface (GUI) that renders the CodaChrome matrix file into an interactive heat map. To generate the matrix file, the PERL scripts retrieve the protein sequences encoded by every fully sequenced bacterial genome recorded in GenBank and construct a BLAST database. Several pre-computed matrix files generated using the 2708 complete bacterial genomes as of 12/3/2013 are available at www.sourceforge.com/p/codachrome. The PERL scripts can be used to generate updated or custom matrix files. When the user selects a "seed organism", each protein encoded in the seed organism is individually queried against the previously generated BLAST database. All statistically significant alignments (E-value < 1e-20) are recorded in a massive list containing hundreds of alignments for each of the thousands of proteins in a typical bacterial proteome. The list of significant alignments is then consolidated and reorganized into a labeled and tab-delimited matrix of best matches. By taking the best BLAST hit, I am intentionally targeting the closest homolog based on sequence identity rather than necessarily attempting to identify the closest ortholog (Altenhoff and Dessimoz 2009). Each column of the matrix corresponds to a protein in the seed proteome, ordered as its corresponding gene is ordered in the seed genome. Each row corresponds to the set of proteins encoded by a specific chromosome or

plasmid represented in the BLAST database. The matrix is populated with the percent identities of pairwise alignments between the seed protein, indicated by the column, and the "best hit" encoded by the plasmid or chromosome indicated by the row. The resulting matrix file is a concise summary of the relationship between the seed proteome and the proteomes encoded by all of the fully sequenced bacterial genomes in GenBank.

Visualization of the CodaChrome matrix file

The data contained in the CodaChrome Matrix File can be visualized using the CodaChrome graphical user interface (GUI) (Figure 3.1).

Figure 3.1 - The CodaChrome graphical user interface



Salmonella enterica 14028S was loaded as the seed organism. The rows were sorted by average proteome identity. The portion of the heat map visualized in the CodaChrome Graph Panel is indicated by the grey box in the Overview Panel. Lighter-than-normal vertical stripes represent large clusters of proteins unique to the seed organism. Dark vertical stripes represent clusters of highly conserved proteins. Buttons embedded in the control panel allow the user to interact with the visualization of the matrix file. A slider at the top of the control panel allows the user to zoom in or out. Replicon arrows and gene selection arrows indicate the alignment selected and described in the information display panel. Finally, the percent identity threshold slider allows users to filter alignments below a specified threshold. For this image, the threshold was set to 20%. The slider also functions as a legend indicating how percent identities are translated into color.

This GUI is programmed in C++/QT and can be compiled to run on most common platforms. It renders the CodaChrome matrix file into a heat map image in which each row corresponds to a replicon in GenBank, each column corresponds to a protein in the seed organism and each pixel is colored according to the percent identity between the two proteins it represents.

The GUI allows users to interact with data represented by this image in many ways. Users may zoom in or out. They may click on the pixels that represent individual alignments to identify which protein-to-protein comparison is represented. They may adjust a threshold to filter out pixels representing low-percentage-identity alignments. They may locate a specific protein or species of interest by typing its name into a dialog box. Finally, they may sort the rows of the matrix by the average percent identity to the seed proteome, to a subset of adjacent seed proteins, or to a single seed protein of interest.

The choice to use C++/QT was motivated by two factors. First, QT is a well developed platform with established cross platform support libraries for GUIs. The write once compile three times was very attractive. Second, it is a mature design platform with WYSIWYG GUI design tools and a custom IDE. Finally, its basis in C++ allows for fast efficient algorithm programming where necessary as well as developed libraries for accessing low level graphics controls necessary for the custom scaling algorithm.

The CodaChrome scaling algorithm

The number of pairwise alignments recorded in a CodaChrome matrix file often exceeds the number of pixels on a typical computer monitor. To visualize the CodaChrome heat map in

its entirety, it must be scaled such that each pixel of the scaled image represents multiple elements of the heat map.

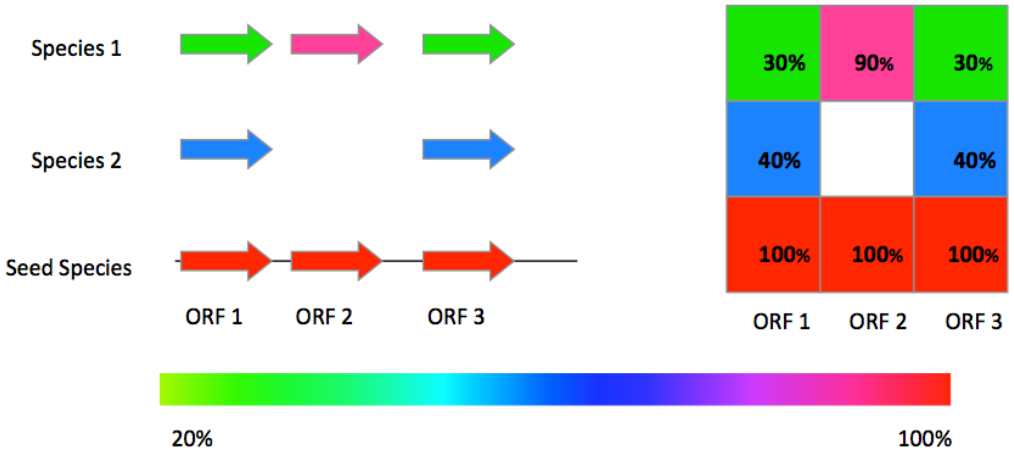
Scaling is accomplished in typical graphics processing algorithms by averaging the values of a block of pixels and coloring a single representative pixel with this value. Alternatively, a single pixel within the block can be chosen at random to represent the whole. Both of these algorithms are unacceptable for CodaChrome because they result in the eclipse of interesting, high value elements by nearby and more common low-value elements. Instead, CodaChrome implements a custom max sampling algorithm that represents a field of pixels with a single pixel of the maximum value within the field. This algorithm causes high-percentage values to "rise to the top" as a user zooms out. This algorithm allows even a single high-percentage value to be visualized while the image is zoomed out by many factors.

Overview Of The Codachrome Graph

CodaChrome generates a BLAST database (Altschul et al. 1990) of all proteins encoded by every finished bacterial genome recorded in GenBank. This database is queried with each protein encoded by a user-selected "seed genome". The pairwise identity between each seed protein and the best match in each organism in the database is indicated by color in an interactive heat map (Figure 3.1). Each column of the heat map corresponds to a protein in the seed organism. The columns are ordered according to the positions of the corresponding genes in the seed genome. Each row of the heat map corresponds to a set of proteins encoded by a chromosome or plasmid in GenBank. The color of each element in the heat map indicates the percent identity shared between the seed protein, indicated by the column, and the most similar protein encoded by the chromosome or plasmid, indicated by the row. When visualized

in this manner, the massive amount of information in the hundreds of thousands of BLAST alignments comparing one proteome to all other proteomes becomes an easily navigable tartan of color, in which patterns emerging above the background reveal important biological relationships and evolutionary events. Because the heat map preserves the gene order of the seed organism, conclusions about genomic events such as insertions and deletions may be inferred even though only protein sequences are compared. The rows of the heat map may be ordered by average percent identity to the entire seed proteome, to a subset of adjacent seed proteins, or to a single seed protein of interest. This method of proteome comparison is represented schematically in Additional File 1: Fig. S1. Note that the patterns in the CodaChrome heat map will differ depending upon which proteome is used as the seed and upon the ordering of the rows. The effects of different parameter choices can be seen by opening two instances of Codachrome and displaying them side-by-side in adjacent windows on the computer screen.

Figure 3.2 - Schematic of the proteome visualization scheme used by CodaChrome



Arrows on the left of the figure represent ORFs. The boxes on the right represent the pairwise identities between the two proteins being compared. Note that the proteins in the seed species are shown in the order in which the genes occur in the genome, but the proteins in species 1 and 2 are not.

In developing CodaChrome, I made certain trade-offs to present the most useful information and the most accurate representation of proteomic relationships. First, the CodaChrome BLAST database is composed only of protein sequences encoded in finished bacterial genomes. The advantage of this method is that users can be confident that gaps or absences in expected patterns represent actual absences of proteins from the organism indicated and not gaps in an incomplete genome assembly. Second, rather than lumping all of the replicons of a given genome together into a single row, CodaChrome assigns each plasmid and chromosome within a genome to its own row. Because the replicons are on separate rows, events such as translocations between replicons or the integration of a plasmid into a chromosome are evident. Third, CodaChrome visualizes the percent identities of amino acid alignments rather than nucleic acid alignments. This choice eliminates the possibility of examining relations between intergenic and other non-coding regions of the genome, but allows CodaChrome to visualize relationships over greater evolutionary distances (States, Gish, and Altschul 1991). Finally, CodaChrome only visualizes the relationships of proteins encoded by the seed genome. Proteins in other organisms that are not found in the seed organism will not be represented in the heat map.

Figure 3.1 illustrates the CodaChrome graphical user interface (GUI) with *Salmonella enterica* 14028S loaded as the seed organism and with the rows ordered by average percent identity to the seed proteome. The protein identity (PID) threshold slider allows the user to adjust a percent identity cutoff to filter information being displayed as well as indicates the color values corresponding to different levels of percent identity: for example, yellow-green corresponds to 20%, blue corresponds to 70% and red corresponds to 100% identity. The

predominantly red and pink band at the bottom of the heat map represents the set of proteomes in which many proteins have nearly 100% identity to the proteins of the seed proteome. Above this band is a sea of green and blue corresponding to proteins from bacteria with more distant relationships to the seed organism. This background of blue and green is lined with several vertical stripes that are unusually dark or unusually light. These stripes correspond to sections of the seed proteome that are especially conserved or especially unique, respectively, in the proteomes in the database.

In the following sections, I show how CodaChrome can be used to address a range of biologically interesting questions. In each case, CodaChrome can provide an answer within a few minutes that previously would have taken considerably more time and/or sophisticated programming. I will also show how anomalous patterns in the CodaChrome heat map can reveal previously unrecognized relationships; behind each of these is an untold biological story.

Interpreting Codachrome Graphs

Identification of the most highly conserved proteins in the bacterial biosphere using CodaChrome

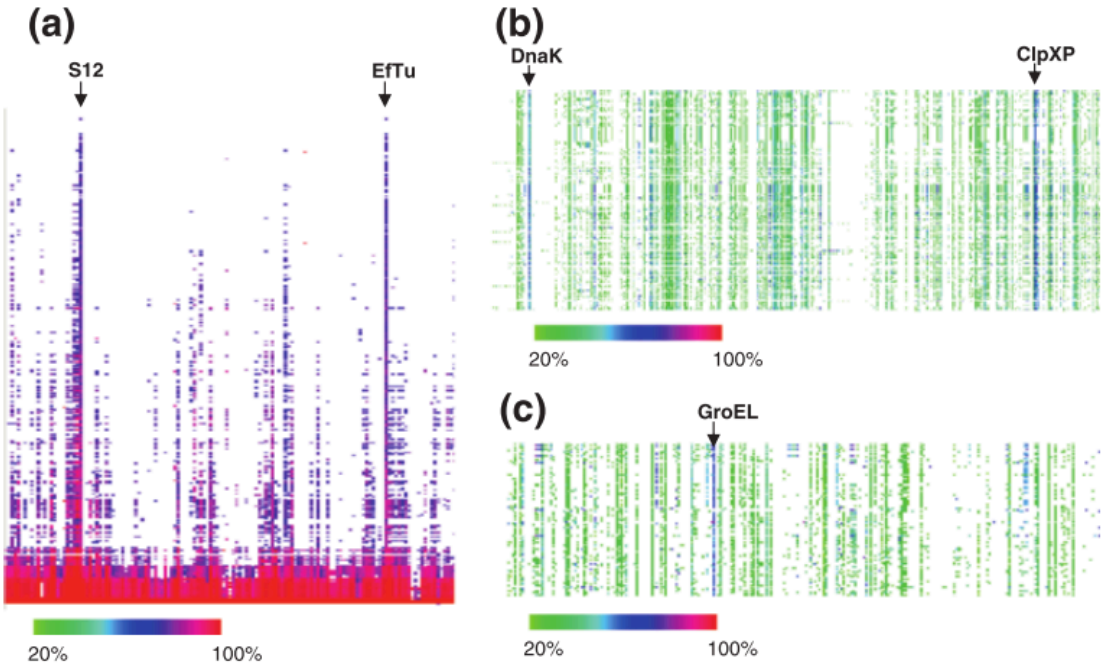
High sequence conservation in proteins from widely divergent organisms is an indication of extreme “purifying” selection; relatively few sequence changes have been tolerated over billions of years. High levels of sequence conservation suggest that the function of the protein is particularly critical for survival, and furthermore that performance of this function constrains its amino acid sequence as a result of specific requirements on correct positioning of residues in

sites involved in ligand recognition, enzymatic activity, and/or interactions with other proteins or macromolecules.

CodaChrome provides an excellent way to identify the most highly conserved proteins in bacteria. After sorting by whole proteome percent identity and applying a percent identity threshold, proteins whose sequences are highly conserved appear as pillars of red and blue that are present even in evolutionarily distant organisms with few other highly conserved homologs.

Two of the most highly conserved proteins in bacteria are elongation factor EF-Tu and the 30S ribosomal subunit protein S12. Figure 3.3a shows that these two proteins are considerably more conserved than the majority of ribosomal proteins, which themselves are among the most highly conserved proteins and account for most of the other pillars in the heat map.

Figure 3.3 - Identification of highly conserved proteins



a) A segment of the CodaChrome heat map with Salmonella enterica subsp. enterica serovar Typhimurium 14028S loaded as the seed sequence and the PID threshold set at 75%. Dark pink and red columns representing pair-wise comparisons of proteins S12 and EF-Tu are indicated. (b) Enlargement of a different region of the same CodaChrome heat map as in the previous panel. PID threshold was set to 20%. The dark column on the right represents two adjacent highly conserved proteins ClpX and ClpP. The dark column on the left represents DnaK. (c) Enlargement of another region of the same CodaChrome heat map as in panel a. PID threshold is set to 20%. The region is centered on the highly conserved protein GroEL.

EF-Tu and S12 are both involved in ensuring the fidelity of protein translation (Ogle and Ramakrishnan 2005). EF-Tu escorts aminoacyl tRNAs to the A-site of the ribosome. If codon-anti-codon pairing is correct, hydrolysis of GTP results in a conformational change that completes the delivery of the charged tRNA into the A-site and causes dissociation of EF-Tu. S12 lies near the A-site at the interface between the two ribosomal subunits and plays a critical role in assessment of correct base-pairing between the codon and anticodon and the subsequent structural rearrangement that occurs when a correct pair is recognized. The high sequence conservation of EF-Tu and S12 can be attributed to both the need to interact with

multiple binding partners and to the intricacy of their functions, which require exquisite sensitivity to the shape of the codon-anticodon pair (Ogle and Ramakrishnan 2005), and apparently place strong constraints on numerous areas of the protein, thus preventing drift due to accumulation of mutations. Strikingly, neither DNA polymerases nor RNA polymerases in this proteome show such stringent conservation, even though high fidelity is also important in replication and transcription.

The CodaChrome heat map shows that the chaperones DnaK (Morano 2007), GroEL (Masters et al. 2009) and the ClpXP protease are also particularly conserved (Figure 3.3b and c). Although these proteins are unrelated, they all interact with a number of "client" proteins, undergo multiple conformational changes, and couple the hydrolysis of ATP to a downstream process that maintains the integrity of proteins in the cytoplasm. For example, ClpXP proteases (Baker and Sauer 2012), which are found widely in bacteria, as well as in mitochondria and chloroplasts, consist of six ClpX monomers in a ring stacked upon two heptameric rings of ClpP. The ATPase subunit of ClpX recognizes degradation tags in proteins (such as the ssrA tag that is added to incomplete proteins released from stalled ribosomes), unfolds the protein and translocates it to the cavity in the rings of ClpP protease. ClpX recognizes over 100 cellular proteins (Flynn et al. 2003; Neher et al. 2006). Mutations in ClpX increase or decrease activity toward certain substrates. Thus it appears that its structure is an evolutionary compromise that allows it to function with a large number of substrates at a reasonable level, even though not optimally with any one (Baker and Sauer 2012). The important functions and complex structures of these multimeric proteins, along with the requirement for interacting with

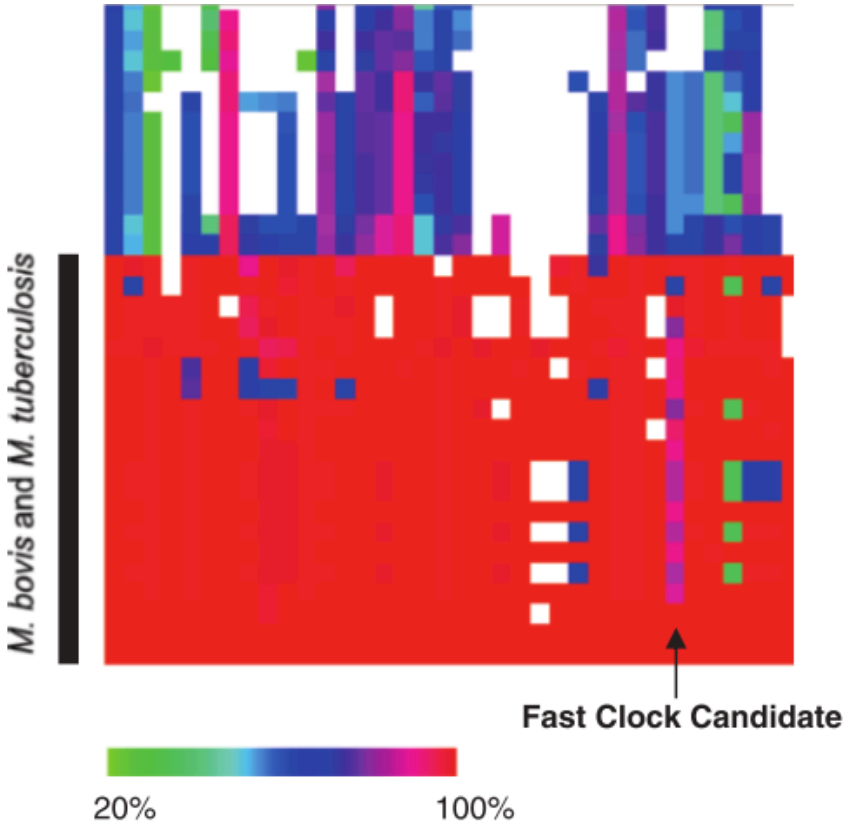
numerous client proteins, have evidently severely constrained the divergence of sequence over vast periods of time.

Identification of Fast-Clock Genes using CodaChrome

Sequencing of 16S rRNA from environmental or clinical samples allows identification of microbes down to the genus and often species level, but often cannot distinguish between strains and closely related species (Kim et al. 2005; Rogall et al. 1990; Santos and Ochman 2004). It is clear that understanding the roles of particular microbes in ecosystems requires finer granularity for classification. Addressing this challenge requires identification of a locus that is moderately conserved across a genus or species of interest but that changes more quickly than the sequence of the 16S ribosomal subunit.

CodaChrome is an ideal tool for identification of "fast-clock" genes within a lineage of interest. The user simply loads the species of interest as the seed, sorts by whole proteome identity, and then looks for pillars of proteins that show marked variation in color. For example, I identified a fast-clock gene useful for the subtyping of several strains of *Mycobacterium tuberculosis* and *Mycobacterium bovis,* all of which possess identical 16S rRNA sequences. *Mycobacterium tuberculosis* H37Rv was loaded as the seed organism, the percent identity threshold was set to 20% and rows were sorted by whole-proteome percent identity. On the bottom of the heat map in Fig. 3, 19 red rows represent the proteins of 19 strains of *Mycobacterium* that are closely related to the seed organism. An ideal fast-clock gene will show significant variation throughout the 19 strains; one such gene, which encodes a PPE family protein, is easily identified by the color variation in the column (Figure 3.4).

Figure 3.4 - Identifying fast clock genes with CodaChrome



Enlargement of a region of a CodaChrome heat map with Mycobacterium tuberculosis H37Rv loaded as the seed sequence and with the PID threshold set to 20%. The rows were sorted by overall percent identity.

A multiple sequence alignment of these sequences confirms the existence of many subtype-specific SNPs and indels. Figure 3.5 contains a matrix of pairwise counts of identical positions, including gaps, in the multiple sequence alignment. Fast-clock genes could be used for rapid identification of strains in clinical or environmental samples based upon differential PCR amplification of variable regions.

Figure 3.5 - Pair-wise percent identities between homologs of PPE34 (YP_177655.1) in closely related strains of Mycobacteria

Percent Identities	F11	GM041182	AF2122/97	Mexico	Tokyo-172	Pasteur-1173P2	CTRI-2	CDC1551	UT205	KZN1435	KZN4207	KZN605	CCDC5079	H37Rv	H37Ra	RGTB423	CCDC5180
F11	100	81	77	78	78	78	92	96	96	96	96	96	50	95	95	95	87
GM041182	81	100	96	92	92	92	77	81	79	77	77	77	31	78	78	78	90
AF2122/97	77	96	100	91	91	91	73	77	75	73	73	73	28	74	74	74	86
Mexico	78	92	91	100	100	100	77	81	79	76	76	76	31	76	76	76	87
Tokyo-172	78	92	91	100	100	100	77	81	79	76	76	76	31	76	76	76	87
Pasteur-1173P2	78	92	91	100	100	100	77	81	79	76	76	76	31	76	76	76	87
CTRI-2	92	77	73	77	77	77	100	92	94	94	94	94	52	89	89	89	84
CDC1551	96	81	77	81	81	81	92	100	98	95	95	95	49	91	91	91	87
UT205	96	79	75	79	79	79	94	98	100	98	98	98	52	93	93	93	87
KZN1435	96	77	73	76	76	76	94	95	98	100	100	100	54	93	93	93	85
KZN4207	96	77	73	76	76	76	94	95	98	100	100	100	54	93	93	93	85
KZN605	96	77	73	76	76	76	94	95	98	100	100	100	54	93	93	93	85
CCDC5079	50	31	28	31	31	31	52	49	52	54	54	54	100	50	50	50	39
H37Rv	95	78	74	76	76	76	89	91	93	93	93	93	50	100	100	100	88
H37Ra	95	78	74	76	76	76	89	91	93	93	93	93	50	100	100	100	88
RGTB423	95	78	74	76	76	76	89	91	93	93	93	93	50	100	100	100	88
CCDC5180	87	90	86	87	87	87	84	87	87	85	85	85	39	88	88	88	100

Red, M. tuberculosis strains; green, M. africanum strains; blue, M. bovis BCG strains.

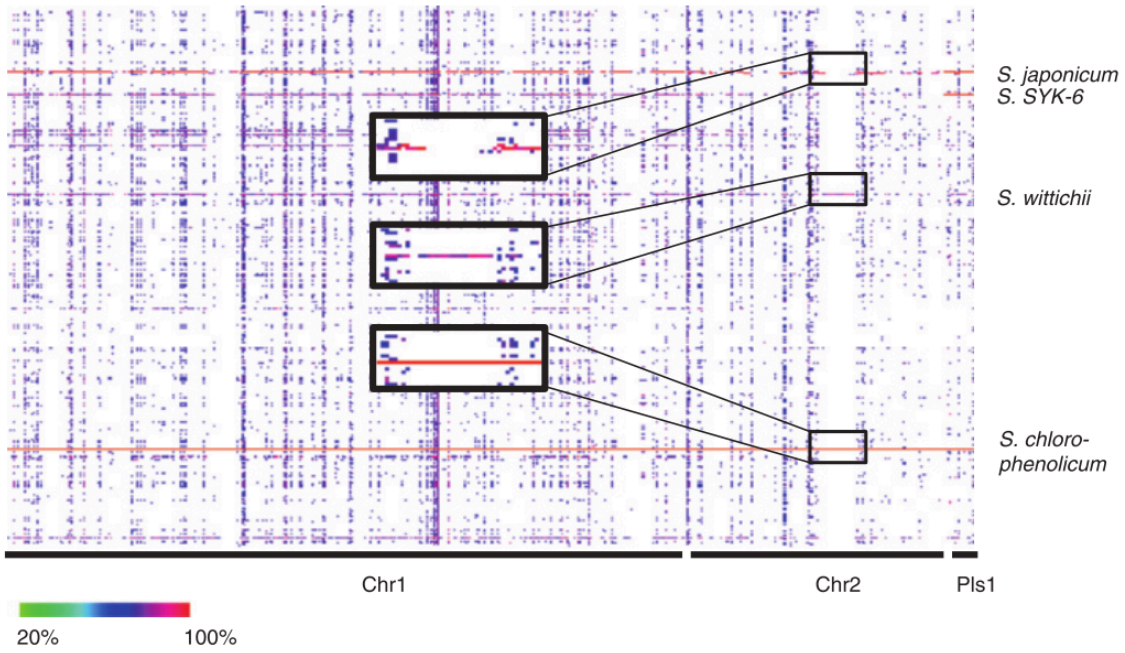
The identification of fast-clock genes by visual inspection of the CodaChrome heat map begs that question of *why* such genes are under selection for unusual variability. Examination of this question can reveal evolutionarily important selective pressures exerted by the environment(s) in which the microbe lives. For example, PPE family proteins in Mycobacteria have been implicated in antigenic variation (Chaitra et al. 2005).

Identification of the evolutionary history of an indel using CodaChrome

Insertions and deletions are common genetic changes that can occur on the scale of single genes or on the scale of hundreds of genes. They are enormously important in

determining characteristics such as virulence, metal tolerance, nutrient requirements, and degradative capabilities. Insertions and deletions are often combined into a single category termed “indels” because in pairwise alignments it is impossible to distinguish whether the observed difference is the result of a deletion from one sequence or an insertion into the other. Additional information from the sequences of more distantly related species is required to resolve the cause of the difference. If the sequence of interest is absent from a more distant relative, parsimony suggests that the indel was absent from the most recent common ancestor as well and arose by insertion in one lineage. Conversely, if the sequence is present in the distant relative, it is more likely that the indel was present in the most recent common ancestor as well and that the indel arose by deletion in the other lineage. The strength of this conclusion can be augmented if there are numerous relatives for comparison. CodaChrome facilitates such analyses by generating a single heat map sufficient for both identification and analysis of indels. Figure 3.6 shows a CodaChrome heat map in which *Sphingobium chlorophenolicum* L-1 was loaded as the seed organism with the percent identity threshold set to 70%.

Figure 3.6 - Investigating the evolutionary history of an indel with CodaChrome



A CodaChrome heat map is shown in which the proteome of Sphingobium chlorophenolicum was loaded as the seed sequence and the rows were not sorted. The boxed regions show a section of the proteome of S. chlorophenolicum that lacks orthologs in the closely related S. japonicum but has orthologs in the more distantly related S. wittichii.

The rows are not sorted, and so remain ordered alphabetically by their GenBank accession numbers. *Sphingobium japonicum* is the most closely related organism included in the database as judged by the red row indicating very high sequence identity extending across nearly all of the heat map. While the majority of proteins in *S. chlorophenolicum* share over 90% identity with homologs in *S. japonicum* (Copley et al. 2012), close comparison reveals several large discrepancies, such as the one in the boxed regions of Figure 3.6. These regions could correspond to a large insertion in the genome of *S. chlorophenolicum* or a large deletion from the genome of *S. japonicum.* The proteome-wide view provided by CodaChrome resolves this question by illustrating that proteins in the boxed segment of the *S. chlorophenolicum*

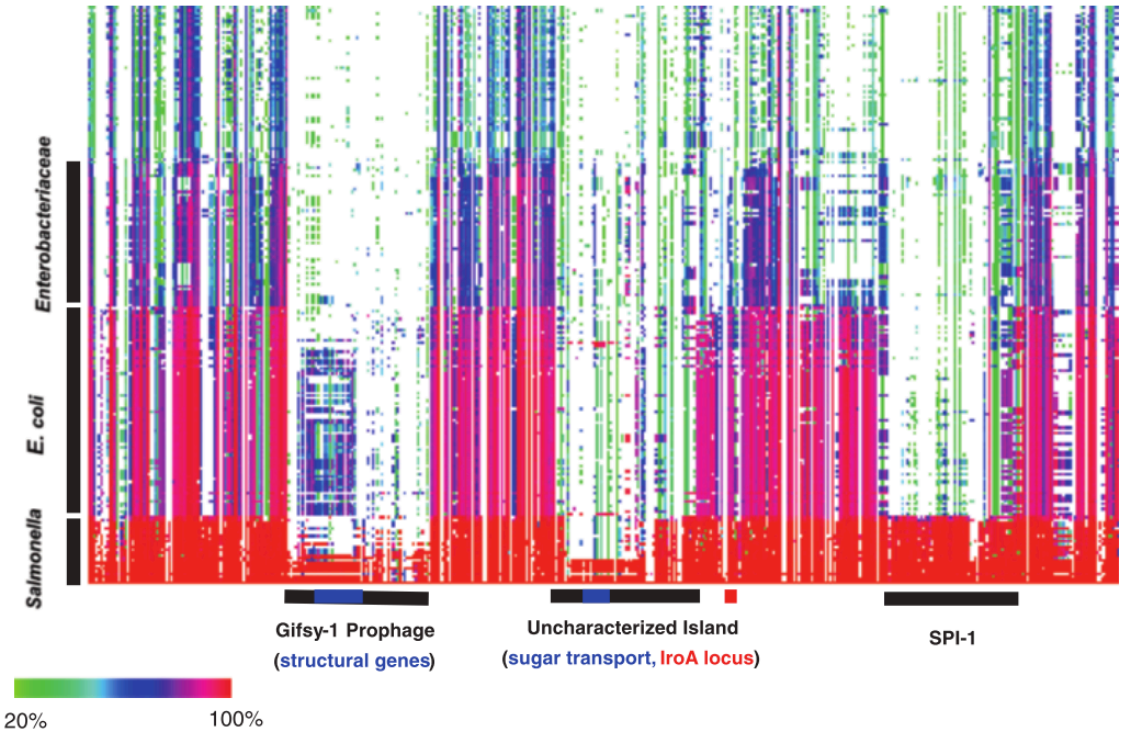
proteome are highly similar to those found in another organism in the database, *Sphingomonas wittichii*. *S. wittichii* is a more distant relative of *S. chlorophenolicum* and *S. japonicum.* Since the proteins included in the boxed region are present in *S. wittichii,* the data support the hypothesis that this segment was present in the common ancestor of *S. wittichii, S. japonicum* and *S. chlorophenolicum* and was lost in the lineage leading to *S. japonicum.*

Examination of questions regarding evolutionary history such as this one are facilitated by CodaChrome because the proteomes encoded by all finished bacterial genomes are simultaneously visualized. Thus, there is no requirement to select, *a priori*, a subset of genomes that is suspected of harboring informative relationships, as is the case when utilizing whole genome alignment software to approach this problem.

CodaChrome facilitates analysis of the pan-genome of bacterial species

CodaChrome is a powerful tool for tracking the movements of blocks of genes that play roles in processes such as virulence, biodegradation, and metabolism. Figure 3.7 shows a region of the CodaChrome heat map in which the proteome of *Salmonella enterica* 14028S was loaded as the seed organism and the rows were sorted by average percent identity shared with *S. enterica* 14028S with the percent identity threshold set at 20%.

Figure 3.7 - Identifying genomic islands in closely related species with CodaChrome



CodaChrome heatmap with Salmonella enterica 14028S loaded as seed organism

The proteomes of the 25 species of *Salmonella* for which finished genome sequences are available sorted to the bottom of the image as a cluster of almost entirely red rows. Above this bright red cluster is a cluster of red and purple rows that represent the proteomes of various strains of *E. coli*. The rows of *E. coli* proteomes are striated with vertical white gaps that represent sections of the *Salmonella enterica* 14028S genome that encode proteins with no clear homology to any *E. coli* proteins. Above the *E. coli* cluster is a cluster of more distantly related Enterobacteriaceae. When sorted this way, large segments that are present in some but not all closely related strains of *Salmonella* are apparent (see horizontal bars).

One of the striking features in the heat map corresponds to the lambdoid prophage Gifsy-1. This prophage is present in about half of the strains of Salmonellae and is largely missing from the *E. coli* cluster aside from a few moderately conserved homologs of the phage

structural proteins. It is entirely missing from the more distantly related cluster of *Enterobacteriaceae* strains. Phage are abundant in nature, and in aquatic habitats they typically out-number microbes by a ratio of 10:1 (Bergh et al. 1989; Rowe et al. 2012; Clasen et al. 2008). Most microbial genomes carry relics of past phage infections in the form of integrated prophages. As clearly revealed in CodaChrome heat maps, these prophages are often the most unique segments of a genome. Under some situations, this characteristic makes them useful loci for strain identification (Helmuth and Schroeter 1994; Young, Ross, and Heuzenroeder 2012). Additionally, some prophages, such as Gifsy-1 in *S. enterica,* have been implicated in pathogenicity (Brüssow, Canchaya, and Hardt 2004). CodaChrome provides a convenient way to visualize the host range of the phage precursor. Simply selecting the cluster of phage proteins and clicking "sort" will reorder the rows such that all replicons that contain similar proteins will sort to the bottom of the heat map.

Prophage insertions are not the only insertions with clinical relevance to leave a clear signature in CodaChrome. Pathogenicity islands also leave a unique pattern. Pathogenicity islands are segments of horizontally transferred genetic material that encode proteins involved in virulence. Pathogenicity islands have been identified in many pathogens, including species of *Escherichia, Vibrio, Listeria, Shigella, Yersina, Salmonella, Pseudomonas, Erwinia, Helicobacter,* and *Agrobacterium* (Sharma et al. 2010). These islands encode proteins involved in a range of processes, including adherence, toxin catabolism, iron uptake, and secretion. Pathogenicity islands range in size from a handful of genes to hundreds. They often have anomalous GC content or codon usage, both hallmarks of horizontal gene transfer. However, when the donor and recipient of the genetic material share similar GC content and codon usage, or when the

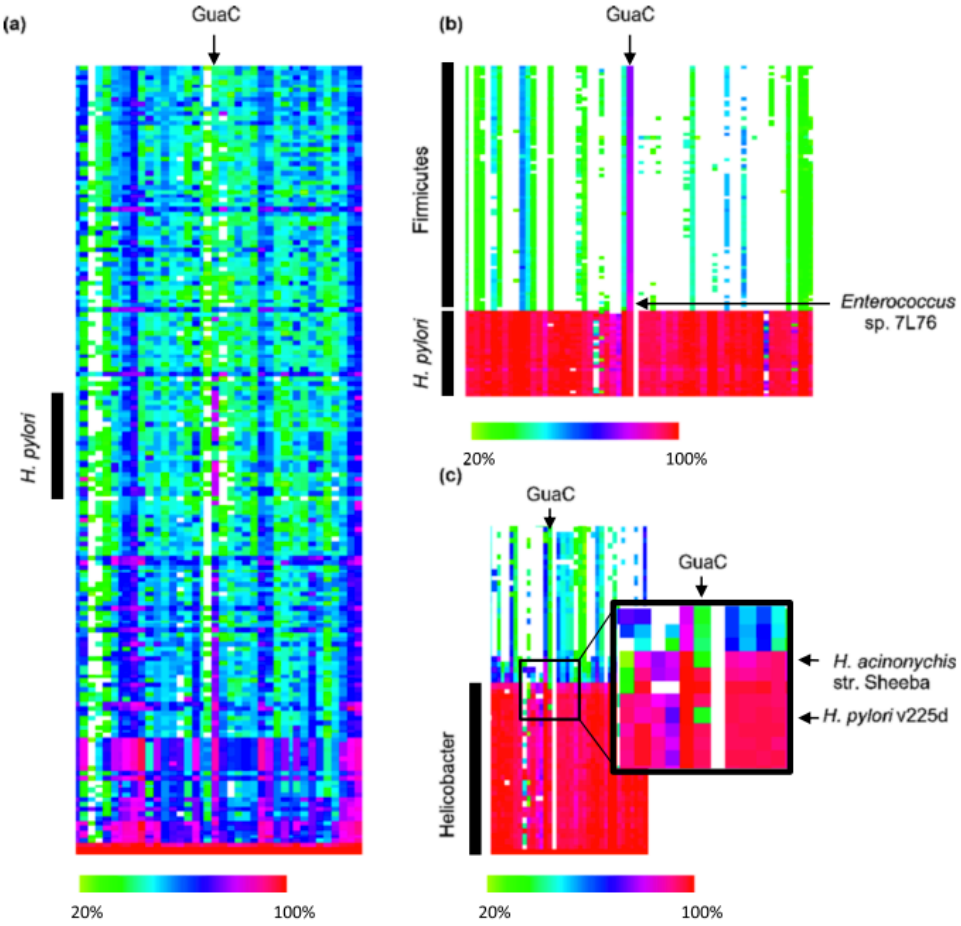
recipient has had sufficient time to equilibrate these properties, these markers may become more difficult to detect or disappear entirely (Ragan, Harlow, and Beiko 2006). The signature of abnormal inheritance that is visualized in CodaChrome remains after these other markers of horizontal gene transfer have been ameliorated. An example of the signature caused by the *Salmonella* Pathogenicity Island 1 (SPI-1) is labeled with a black bar in Figure 3.7. SPI-1 encodes 47 proteins, most of which are involved in assembly or function of the Type III secretion system (Zhou and Galán 2001), or are effector proteins that are injected into host cells through the needle of the secretion complex.

Like the prophage on the left of Figure 3.7, SPI-1 is conspicuously lacking from the closely related cluster of *E. coli* as well as the cluster of enterobacteriaceae. Unlike the prophage, however, SPI-I is present in all sequenced strains of *Salmonella*, reflecting its essential role in *S. enterica* pathogenesis. Figure 3.7 also shows a third, largely uncharacterized region of anomalous sequence conservation. This region encompasses an odd collection of genes, including genes involved in iron metabolism, sugar transport and many conserved proteins of unknown function.

Use of CodaChrome as a discovery tool

The previous examples have illustrated the use of CodaChrome for analysis of biologically interesting questions. However, CodaChrome can also be used as a discovery tool due to its ability to reveal patterns in sequence conservation that are easily recognized by the human eye. These striking patterns indicate relationships, or in some cases, the absence of expected relationships, that might not have been previously recognized. An example is shown in Figure 3.8a.

Figure 3.8 - CodaChrome heat maps reveal unexpected sequence relationships



a) A region of the CodaChrome heat map in which B. subtilus was loaded as the seed organism, the rows were sorted according to whole proteome identity, and alignments with pairwise percent identities below 20% were filtered out. b) A region of the CodaChrome heat map in which H. pylori P12 was loaded as the seed organism, the rows were sorted according by identity to GuaC of H. pylori P12, and alignments with pairwise percent identities below 20% were filtered out. c) Enlargement of the region around GuaC of a CodaChrome heat map with H. pylori P12 loaded as the seed proteome and rows sorted by whole proteome percent identity.

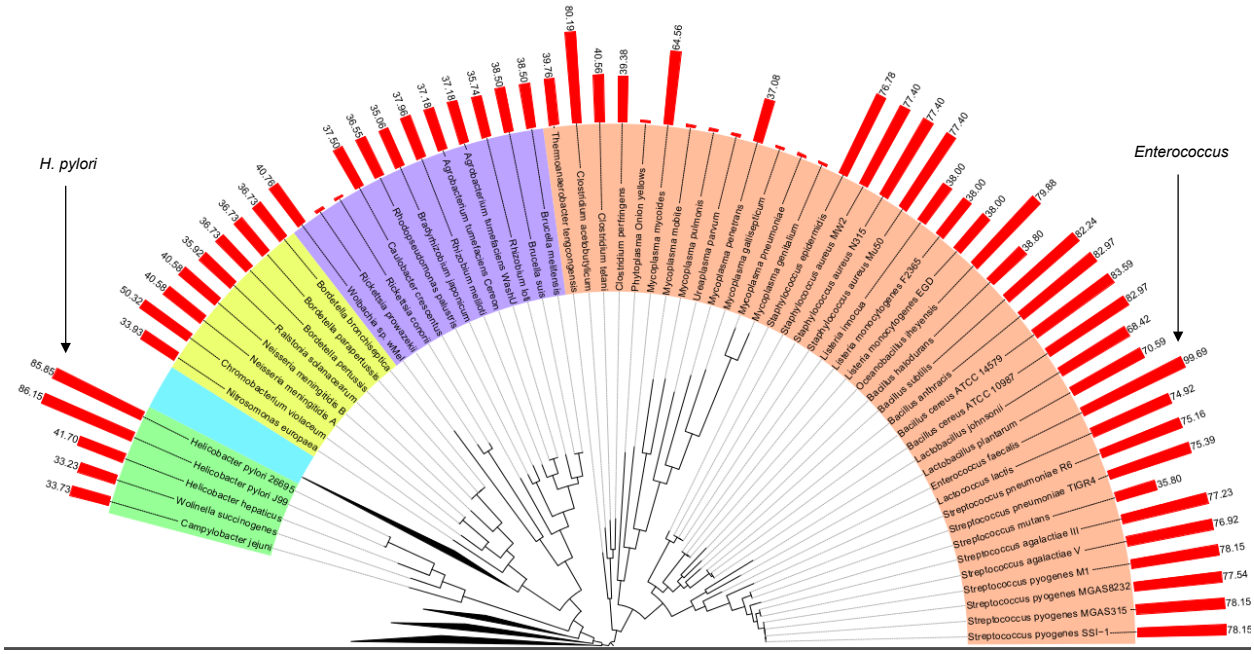
When the proteome of the Firmicute *Bacillus subtilus* was used as the seed in CodaChrome and replicons were sorted by average percent identity to the entire seed proteome with a percent identity cutoff of 20%, I noticed a vertical stripe of red and pink

standing out in a region far above the horizontal band of red and pink corresponding to proteomes closely related to that of *B. subtilis* (Figure 3.8a). This vertical stripe is due to unexpectedly high sequence identity between *B. subtilis* guanosine reductase (GuaC) and homologous proteins in several strains of the Proteobacterium *Helicobacter pylori.*

Guanosine reductase is involved in purine salvage. It converts guanosine monophosphate to inosine monophosphate (IMP), which in turn can be converted to adenosine monophosphate. Since *H. pylori* lacks the pathway for *de novo* synthesis of inosine monophosphate, its ability to salvage purines from its environment is essential for growth (Liechti and Goldberg 2012). GuaC from *B. subtilis* is 81% identical to GuaC from *H. pylori*; this is more than twice the level of percent identity shared by most homologous proteins in these bacteria as assessed by the color of the rows corresponding to *H. pylori* in the CodaChrome graph in which *B. subtilis* is the seed organism. Not even highly conserved proteins such as S12 and EF-Tu share such a high percent identity; S12 and EF-Tu in the two bacteria are 68% and 74% identical, respectively.

I investigated this surprising relationship further by loading the proteome of *H. pylori* as the seed sequence (Figure 3.8b) and sorting the heat map by percent identity to *H. pylori* GuaC protein with a PID cutoff of 20%. This view shows that *H. pylori* GuaC has closely related homologs in many Firmicutes; the closest homolog is found in *Enterococcus sp.* 7L76. Figure 3.9 shows values for the percent identity between GuaC from *Enterococcus sp.* 7L76 and its closest homolog in other microbes represented on a bacterial phylogenetic tree (Letunic and Bork 2011). Indeed, *Enterococcus sp.* 7L76 GuaC is more closely related to homologs in *H. pylori* than to homologs in many closely related species of Firmicutes.

Figure 3.9 - Percent identity between GuaC from Enterococcus sp. 7L76 and closest homolog



Mapping of the pair-wise sequence identities between GuaC from Enterococcus faecalis and the closest homolog in other representative bacteria onto a phylogenetic tree.

From this pattern in CodaChrome, I infer that the gene encoding GuaC was transferred from a recent ancestor of *Enterococcus sp.* 7L76 to a recent ancestor of *H. pylori*. *H. pylori* colonizes the stomach in humans, while many species of *Enterococcus* are found lower in the intestinal tract (Sekirov et al. 2010). *H. pylori* might have acquired a gene encoding GuaC from lysed enterococci that had been ingested through food or fecal-oral contamination and killed by the highly acidic environment of the stomach. The acquisition of GuaC, as well as other enzymes involved in purine salvage, may have allowed this bacterium to lose the genes required for *de novo* purine synthesis as it adapted for growth within the nutrient-rich environment of the stomach.

Re-sorting the CodaChrome heat map by percent identity across the entire proteome reveals another surprising finding (Figure 3.8c). The closest homologs to *H. pylori* GuaC are

unexpectedly distant in two closely related species of *Helicobacter*; *H. acinonychis* str. Sheeba, a gastric pathogen of big cats that is believed to have jumped from humans to large felines about 200,000 years ago (Eppinger et al. 2006), and *H. pylori* v225d, a clinical strain isolated from a Piaroa Amerindian with acute gastritis (Mane et al. 2010). The proteins corresponding to the green blocks indicated by arrows in Figure 3.8c are annotated as IMP dehydrogenases. These annotations are undoubtedly correct, as these proteins are >97% identical to IMP dehydrogenases from other strains of *H. pylori.* Thus, the mechanisms by which these bacteria salvage purines for synthesis of nucleotides are unknown.

Future Work

Installing the backend of CodaChrome requires a fair amount of work. A future web based version would make the program much more accessible. Similarly, a user friendly method for allowing users to upload custom genomes or search against custom subsets of genomes would be useful. Finally, incorporating the known tree of life hierarchy or an ad hoc clustering of the species allowing the collapse or exclusive viewing of clades of interest would be a very useful tool. While syntenic information is of interest, clustering along the x axis would allow genes spread throughout the genome with similar conservation patterns to be highlighted potentially providing valuable biological insights.

Conclusion

CodaChrome is powerful enough to visualize relationships between thousands of bacterial genomes, yet sensitive enough to visualize the deletion of a single protein from a single species. CodaChrome builds users' intuition about the relationships between genomes.

It establishes what is “normal” and in doing so provides a baseline for identifying what is unusual about proteins within a particular genome. Unlike many visualization tools in which data are used to render a static image, CodaChrome is interactive. Patterns are not just visualized, but can easily be investigated in more detail by sorting the rows in various ways, zooming in on a region of interest, and/or mousing over individual blocks to obtain specific information about the pairwise relationship represented by the block.

Sorting the CodaChrome heat map by average percent identity across the entire proteome provides users with an empirical estimate of the phylogenetic distances between the proteome of the seed organism and the proteomes of all other organisms in the database. Sorting by subsets of the seed proteome allows users to investigate individual proteins or clusters of proteins that have experienced atypical selection pressures or horizontal gene transfer.

As demonstrated in the examples above, this tool allows investigation of a broad range of biological questions at scales ranging from individual proteins to groups of proteins transferred *en bloc* to the entire proteome. Further, CodaChrome visualizations contain more information than the individual alignments of which they are composed. Visualizing all of the alignments simultaneously forms patterns indicative of biological phenomena that could not easily be identified by examining the alignments individually.

CodaChrome operates in terms of the tree of life. The tree of life is a hierarchical network that connects every living organism. It is rooted somewhere between Archea and Bacteria where we think the LUCA would connect up if we knew its genome sequence. This rooting is arbitrary though. CodaChrome effectively reroots the tree of life at the organism of

interest and visualize the conservation decay of each gene as one moves outward. The CodaChrome graph is the equivalent to peering down into the tree of life from one particular leaf rather than from the bottom up. The most conserved genes reach back into more and more distantly related organisms. Horizontal gene transfer events appear as discontinuities in these long running streaks; They are islands of conservation at odds with the global conservation of other genes.

The greatest strength of CodaChrome is its ability to contextualize protein conservations. If a protein shares 70% identity with a protein in another organism. Is that high? Is that low? Is it part of a similarly conserved operon? Questions like these are instantly obviously with CodaChrome. With the whole picture available, I am instantly able to identify what is normal and what is exceptional. CodaChrome builds an intuition about the relationships between different genomes.

Unfortunately, in the case of *S. chlorophenolicum*, there are only a handful of fully sequenced relatives close enough to be informative limiting the utility of CodaChrome. As the cost of genome sequencing decreases and the number of sequenced genomes increases, so will the power of CodaChrome mediated comparative genomics.

Chapter 4 Early transcriptional response of *S. chlorophenolicum* to PCP stress

Introduction

In this chapter I show that, in addition to the known PCP induced degradative enzymes, many other genes are induced in response to PCP stress. Some of these are likely induced promiscuously and have an endogenous role responding to stimulus by a molecule related to PCP.

PCP and its degradation intermediates stress a cell in many ways likely triggering multiple stress responses. Additionally, genome sequencing revealed that the *S. chlorophenolicum* genome, and especially the second chromosome, encodes many genes involved in putative degradation pathways (Copley et al. 2012). Promiscuous transcriptional activation has been postulated as an important step in the evolution of new pathways (van der Meer 1997). Finally, the PCP degradation pathway was worked out through very targeted methods and many involved genes may have been missed. These genes may play important roles in PCP degradation or provide clues to the ancestral functions of the PCP degradation enzymes. For these reasons, I expected the transcriptional response of *S. chlorophenolicum* to PCP stress to be rich and informative. I biased the response toward primary transcriptional effects by collecting samples 15 minutes after induction with PCP. This time point is long enough to elicit a robust transcriptional response but too short for new protein to be synthesized from these transcripts, generating secondary transcriptional effects. Samples were

grown to mid log phase, induced with 100uM PCP, and samples were collected immediately before induction and 15 minutes after induction as described in the Methods.

Results Of Sequencing

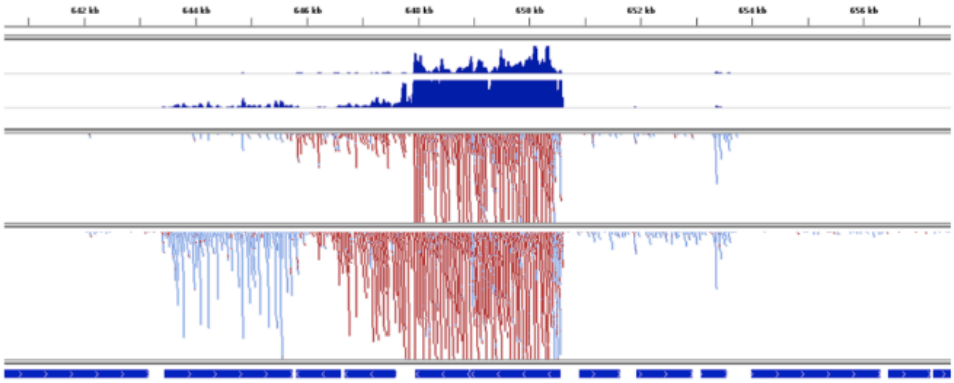
After preparing the library as described in Methods, we ran the library on a single lane of the Illumina Hi-Seq on a 100bp single end run. The machine generated about 120 million reads. Since the library contained two multiplexed samples, the reads generated represented 60 million reads per sample.

After performing quality control checks on the read data, much of the data was unusable. Read quality deteriorated rapidly approaching the 3' end of the read, such that only the first 30 base pairs of the 100bp reads were of high quality. After truncating the reads to the high quality first 30 base pairs, only ~20% mapped to the *S. chlorophenolicum* genome. Of the 20% that mapped, only about 20% of reads were non-ribosomal. Of the 2x60 million reads, only 2x12 million mapped and only 2x2 million mapped and were non-ribosomal. Two million informative reads distributed over 4000 genes in the *S. chlorophenolicum* genome yielded 500 reads per transcript per sample on average.

After performing the quality control filters, the mapped reads were imaged in IGV (Robinson et al. 2011). Scanning through various loci, despite the heavy cuts from the quality control filters, the reads that remain are consistent with high quality library preparation and sequencing. First, reads predominantly originate from annotated coding regions of the genome. Second, the library was generated in a strand specific fashion and the high quality mapped reads all match the orientation of the ORF they encode. Finally, the genes that are known to be

induced in the presence of PCP are all up-regulated as expected. These three positive controls for quality are all visible in the IGV plot of the *pcpBD*, *pcpR* loci pictured in Figure 4.1.

Figure 4.1 - IGV plot of the pcpR, pcpD, pcpB locus



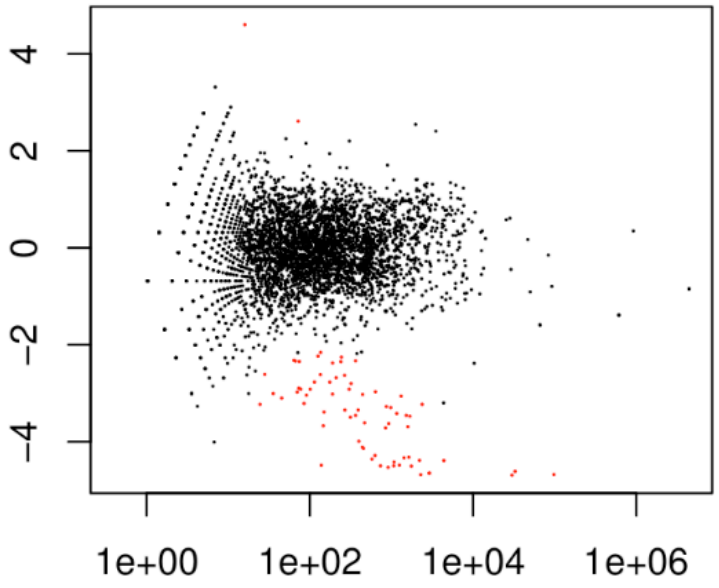
The top track indicates the base pair location in the genome of the locus imaged in IGV. The next two tracks in blue plot the average coverage before and after PCP stress. In the next two tracks individual reads are plotted and colored red or blue if they originate from the forward or reverse strand. The final track plots the open reading frames in this region of the genome with their direction indicated by the arrow. The three open reading frames dominated by red mapped reads from right to left correspond to pcpB, pcpD and pcpR.

Global Transcriptional Response

After performing quality control, I investigated the global transcriptional response of the *S. chlorophenolicum* transcriptome to PCP stress. In addition to the known PCP catabolizing enzymes, several other genes and operons were also found to respond to PCP stress.

To identify significantly differentially expressed genes I preformed a DE-SEQ analysis (Anders and Huber 2010) across the two conditions tested. The results of this analysis are displayed in Figure 4.2 with an ME plot showing significantly differentially expressed genes colored red.

Figure 4.2 - DE-seq ME plot (fold change vs read depth) before and after PCP stress

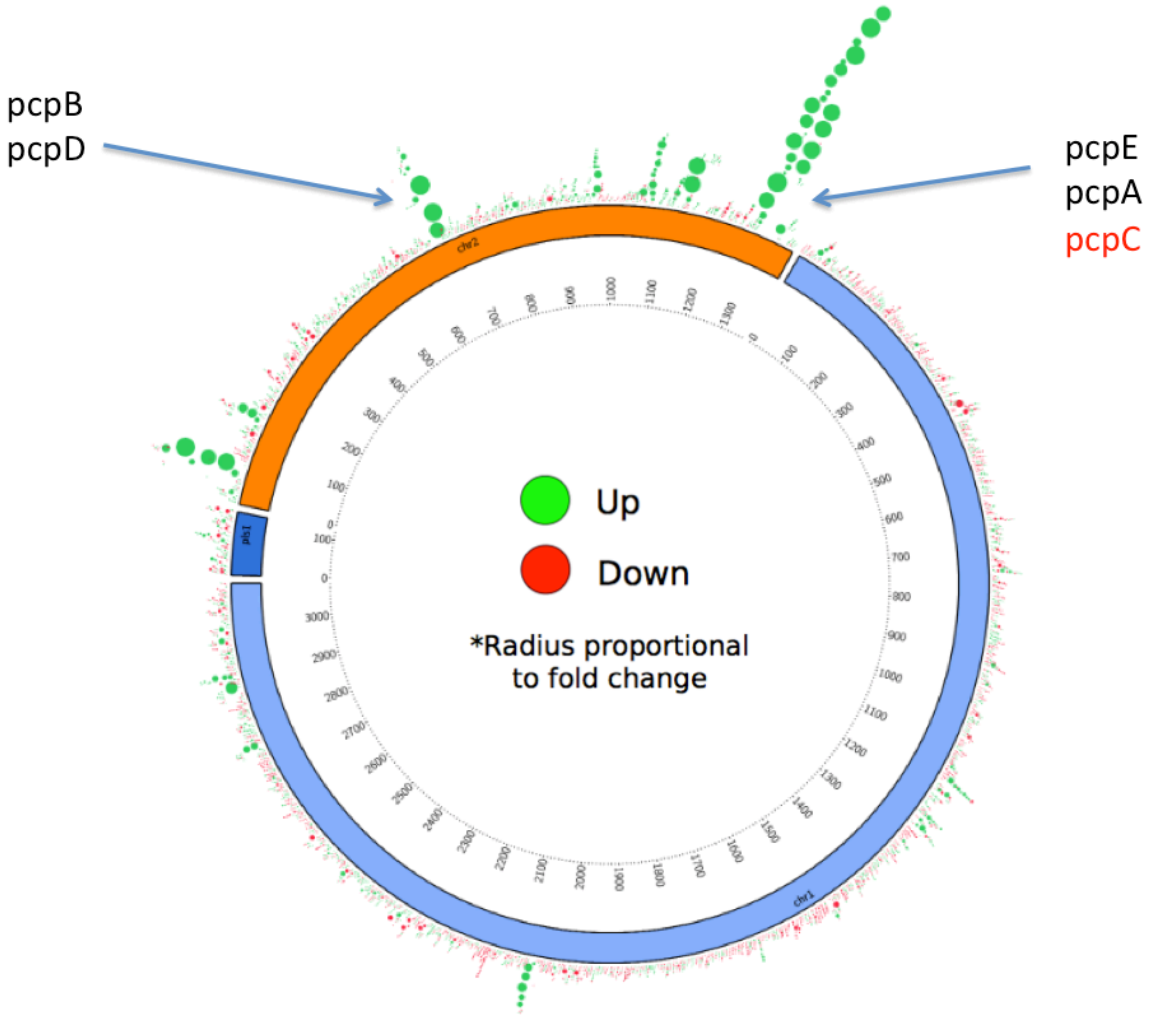


Each dot represents a gene. The X axis is DE-seq normalized read count. The Y-axis is log fold change of gene expression just before 100uM PCP stress over gene expression 15 minutes after the stress. Points representing significantly differentially expressed genes are colored red.

From this graph, I can see that the vast majority of significantly differentially expressed genes are up-regulated after the addition of PCP. These include the known PCP genes as well as many genes that were previously unknown to be regulated during PCP stress.

When visualizing differential expression in an ME plot, all gene order information is lost. Gene order is very informative in bacteria where genes are often ordered into operons or adjacent functionally related clusters. I implemented a method to visualize the global transcriptional response of *S. chlorophenolicum* using the program Circos (Krzywinski et al. 2009) that would preserve gene order information. When visualized in this way many co-regulated clusters of genes are apparent, many of which are organized into operons (See Figure 4.3).

Figure 4.3 - Global gene expression change in S. chlorophenolicum



Inner circle is the kb position of replicon indicated by the colored segment in the next circle moving outwards. The radius of the red and green circles is proportional to the fold change of the gene they represent. Colored circles are arranged so that they are over the kb position of the gene they represent. Circles are stacked when crowding necessitates it to keep the circle over its position in the genome.

Many genes were differentially regulated in response to PCP. The majority of these up-regulated genes were located on chromosome 2 consistent with the observation that many of the putative degradative pathways were identified on this chromosome. Promiscuous

activation appears rampant. Five clusters of contiguous upregulated genes and there annotations are described in Figure 4.4.

Figure 4.4 - Contiguous up-regulated genes

IMG ID	Annotation	Before	After	FC	P-value
2503962996	"drug resistance transporter, EmrB/QacA subfamily"	199	2125	11	2.00E-02
2503962997	"efflux pump membrane protein (multidrug resistance protein A)"	254	2788	11	2.00E-02
2503962998	"efflux transporter, outer membrane factor (OMF) lipoprotein, NodT family"	139	1713	12	7.00E-03
2503962999	"transcriptional regulator, TetR family"	42	338	8	4.00E-03
2503963000	"Response regulators CheY-receiver and winged-helix DNA-binding domain"	16	124	8	6.00E-03
2503964000	“related flavin-dependent oxidoreductases"	181	1782	10	2.00E-02
2503964001	"Esterase/lipase"	71	536	8	1.00E-02
2503964002	"Arabinose efflux permease"	148	3362	**23**	2.00E-03
2503964003	"Conserved protein/domain flavoprotein oxygenases, DIM6/NTAB family(EC:1.14.13.3)"	54	1100	**21**	5.00E-04
2503964004	pcpX - "Outer membrane receptor proteins, mostly Fe transport"	222	5557	**25**	3.00E-03
2503964005	"Acyl-CoA synthetases (AMP-forming)/AMP-acid ligases II(EC:6.2.1.3)"	61	667	11	2.00E-03
2503964502	pcpY - "Outer membrane receptor proteins, mostly Fe transport"	156	3118	**20**	3.00E-03
2503964503	"Glycosyl transferase family 2."	230	402	2	1.00E+00
2503964504	**pcpR - "Transcriptional regulator"**	282	2348	8	6.00E-02
2503964505	**pcpD - "Flavodoxin reductases (ferredoxin-NADPH reductases) family 1(EC:1.14.13.82)"**	2580	63031	**24**	6.00E-02
2503964506	**pcpB - "2-polyprenyl-6-methoxyphenol hydroxylase FAD-dependent oxidoreductases"**	7372	188476	**26**	8.00E-02
2503964789	"efflux transporter, outer membrane factor (OMF) lipoprotein, NodT family"	79	402	5	6.00E-02
2503964790	"The (Largely Gram-negative Bacterial) Hydrophobe/Amphiphile Efflux-1 (HAE1) Family"	163	1573	10	2.00E-02
2503964791	"RND family efflux transporter, MFP subunit"	70	714	10	3.00E-03
2503964792	"hypothetical protein"	1	4	5	1.00E+00
2503964793	"Response regulators consisting of a CheY-like receiver domain and a winged-helix DNA-binding domain"	159	704	4	2.00E-01
2503964794	"Signal transduction histidine kinase(EC:2.7.13.3)"	121	607	5	9.00E-02
2503965065	"aminodeoxychorismate synthase, component I, bacterial clade(EC:2.6.1.85)"	8	63	8	1.00E-02

2503965066	"hypothetical protein"	202	4219	**21**	4.00E-03
2503965067	"Glutathione S-transferase(EC:2.5.1.18)"	120	1574	13	5.00E-03
2503965068	pcpZ - "Outer membrane cobalamin receptor protein"	96	2059	**21**	1.00E-03
2503965069	"Predicted esterase"	49	861	18	5.00E-04
2503965070	**pcpE - "maleylacetate reductase"**	400	8377	**21**	1.00E-02
2503965071	"formyltetrahydrofolate deformylase (EC 3.5.1.10)(EC:3.5.1.10)"	17	153	9	2.00E-03
2503965072	"methenyltetrahydrofolate cyclohydrolase (EC 3.5.4.9)"	18	137	8	6.00E-03
2503965073	**pcpM - "Transcriptional regulator"**	29	199	7	1.00E-02
2503965074	**pcpA - "Glyoxalase/Bleomycin resistance protein/Dioxygenase superfamily."**	2242	57846	**26**	5.00E-02
2503965075	"Kef-type K+ transport systems, membrane components"	47	752	16	6.00E-04
2503965076	"Metal-dependent hydrolases of the beta-lactamase superfamily III"	57	364	6	2.00E-02
2503965077	"Outer membrane receptor proteins, mostly Fe transport"	136	619	5	2.00E-01
2503965078	**pcpC - "Glutathione S-transferase"**	3329	17350	5	1.00E+00
2503965079	"Predicted glutathione S-transferase"	278	3092	11	2.00E-02
2503965080	"LysR Transcriptional regulator"	48	832	17	5.00E-04

2503965092	"PilZ domain./Helix-turn-helix."	62	1198	19	6.00E-04
2503965093	"Uncharacterized protein conserved in bacteria"	226	5646	**25**	3.00E-03
2503965094	"AraC-type DNA-binding domain-containing proteins"	514	605	1	1.00E+00
2503965095	"Protocatechuate 3,4-dioxygenase beta subunit(EC:1.13.11.1)"	172	4408	**26**	2.00E-03
2503965096	"Uncharacterized protein conserved in bacteria"	76	1747	**23**	7.00E-04
2503965097	"Glutathione S-transferase(EC:2.5.1.18)"	107	2398	**22**	1.00E-03
2503965098	"Glutathione S-transferase"	91	2047	**23**	1.00E-03
2503965099	"Nitroreductase(EC:1.-)"	135	2717	**20**	2.00E-03
2503965100	"Outer membrane receptor proteins, mostly Fe transport"	229	2952	13	1.00E-02

Among the likely candidates of promiscuous activation is a large cluster of nine genes involved in a putative nitroaromatic degradation pathway located near the cluster of genes encompassing *pcpC, pcpA, pcpM,* and *pcpE*. The promoters of this cluster do not contain the PcpR motif and are likely transcribed as a result of promiscuous activation by the adjacent regulators.

It is fascinating to note that a paralogs of *pcpE* is significantly up-regulated in response to PCP. The homolog is part of a putative operon encompassing three genes including a putative monooxygenase. The putative operon is oriented divergently from a putative LTTR. The intergenic region between the LTTR and the first gene of the operon does not contain any instances of the ATTC-N_7-GAAT and is likely induced by promiscuous activity of the adjacent LTTR rather than by PcpR.

On chromosome 1, two operons encoding efflux pumps are significantly up-regulated. These genes may play a role in excluding PCP or its breakdown intermediates from the cell.

Phage Shock Response

Homologs of the *E.* coli phage shock response proteins are encoded in the *S. chlorophenolicum* genome. Because this response is activated when the bacterial proton gradient is destabilized I expected activation in the presence of PCP. However, these genes were not highly induced. A moderate two fold increase in transcript levels was observed.

Conclusion

There is a rich immediate transcriptional response to PCP in *S. chlorophenolicum*. In addition to the known PCP degradation genes, I also observed the induction of a diverse set of genes involved in functions ranging from transport and efflux to the degradation of related aromatic compounds.

Chapter 5 Expanded Roles for PcpR and PcpM

Introduction

Three facts were known about the regulation of the pentachlorophenol degradation genes when I started this project:

1. Two genes encoding LysR type transcriptional regulators (LTTRs) are located near the genes encoding the PCP degradation enzymes: *pcpR* and *pcpM* (Cai and Xun 2002)
2. A knockout of *pcpR* fails to induce any of the PCP degradation genes in the presence of PCP and fails to degrade PCP (Cai and Xun 2002).
3. A knockout of *pcpM* still induces all the PCP degradation genes in the presence of PCP and is entirely competent at degrading PCP (Cai and Xun 2002).

Cai et al. also noted a 5 fold increase in PcpA enzymatic activity in the *pcpM* knockout. They attempted to find a corresponding change in *pcpA* gene expression but the RT-PCR method they used was not quantitative and they were unable to. Based on these observations, Cai et al. suggested that PcpR is the master "PCP regulator" and that PcpM is not critical for PCP degradation. Yet, when I compare the ratio of synonymous mutations to non-synonymous mutations between PcpM and its ortholog in *Novosphingobium resinovorum,* I see a strong signature of positive selection suggesting a functional role for PcpM in the cell. Furthermore, the ortholog of *pcpM* is maintained divergently oriented from the ortholog of *pcpA* despite the fact that synteny is largely shuffled between these two species. Given these observations, what is the role for PcpM in the cell?

The role of PcpR is more clearly worked out but many questions remain. What other targets might PcpR regulate in the cell? To answer these questions, in this section, I leverage the genome sequence and global transcriptional response of *S. chlorophenolicum* to uncover an expanded PCP regulatory motif and characterize the interaction between PcpR and this expanded motif. I uncover new genes regulated by PcpR and show that the novel genes are not essential for PCP degradation. Finally, I present evidence that PcpM is transcribed and translated and potentially interfering with PcpR binding at the PcpA promoter.

Bioinformatic Experiments

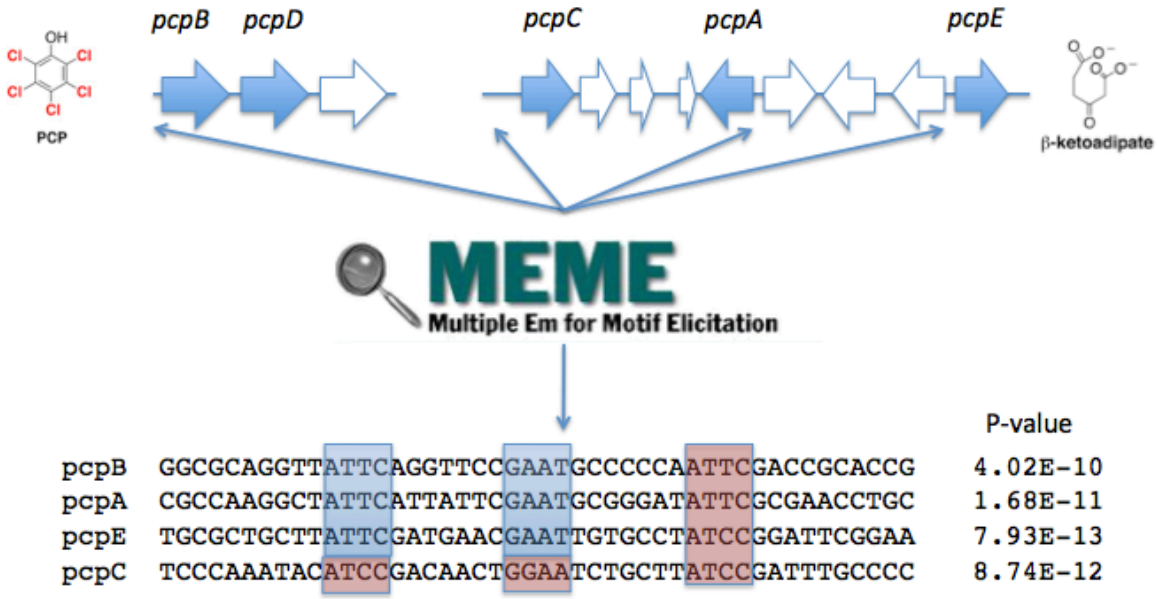
Identifying an Expanded Motif

A putative PCP regulating LysR motif was previously identified by going over the intergenic sequences upstream of the known PCP genes by hand and looking for sequences that fit the NTNN-N_7-NNAN consensus motif of an LTTR. An ATTC-N_7-GAAT motif was found upstream of *pcpBD*, *pcpA* and *pcpE*. Additionally, a different more degenerate pattern was found upstream of *pcpC* .

I undertook replicating this approach computationally in an unbiased manner. A computation approach has several advantages. The background nucleotide frequencies and the size of the intergenic region examined can be taken into account. Also, if other motifs are over represented in the sequences examined they will be found as well regardless of their form. The strategy I used was to take the complete intergenic sequences upstream of the known PCP induced proteins and scan them through MEME, a program that uses an expectation maximization algorithm to identify over represented motifs (see Figure 5.1). I used the average

nucleotide frequencies of the intergenic regions as the background frequency. This is important because the GC content of the intergenic regions and the genome as a whole differs greatly

Figure 5.1 - Identifying the PCP LysR Type Motif



The intergenic regions upstream of the genes responsible for PCP degradation were examined for overrepresented motifs. A single motif was found in every promoter. This motif includes the previously discovered motif (blue) as well as previously unrecognized motif (red).

In addition to the previously identified motifs (colored blue) an additional inverted repeat located downstream was identified. Also, a new putative lysR motif was found upstream of *pcpC*.

New targets of PcpR

The expansion of the motif from 8 to 12 base pairs is significant for two reasons. First, it informs the mechanism of PcpR activation. The typical LysR binding motif contains a single perfect inverted repeat of the consensus sequence separated by 7 base pairs. The fact that all of the induced genes have three perfect inverted repeats suggests that PcpR behaves

differently than the canonical LTTR. Perhaps it has poor affinity to DNA and the third site is necessary for sufficient binding affinity. Perhaps, to generate the extreme induction of its target genes, PcpR has to generate an extreme bend in the DNA requiring strong affinity to the DNA. I test these ideas with gel shifts of variously mutated DNA later in the chapter.

The second reason this expanded motif is significant is that a 12 base sequence is relatively unique. While I would expect an 8 base sequence to appear many times randomly, a 12 base sequence is much less common. This means that I can scan through the genome in an unbiased manner and when instances of the motif are uncovered in intergenic areas, it is highly probable that they are non-random. I scanned the genome for instances of the 12 base pair motif and cross-referenced it to the RNAseq data. All of the most significant intergenic instances of the motif on Chromosome 2 were over expressed by almost 20 fold. This list recovered all of the known PCP induced genes, which is unsurprising given they were used to generate the motif in the first place, as well as four new candidate genes (see Figure 5.2). One of these genes was *pcpC.* Known to participate in the degradation of PCP, the inclusion of *pcpC* on this list was surprising as it was not known to be regulated in response to PCP and thought to be simply constitutively expressed. The other three genes were all annotated as transporters. I designated them *pcpX*, *pcpY,* and *pcpZ.*

Figure 5.2 - Putative targets of PcpR

matched sequence	Fold Change	Predicted Function
ATTCAGGTTCCGAATGCCCCCAATTCG	26	**pcpB** - pentachlorophenol monooxygenase
ATTCATTATTCGAATGCGGGATATTCG	7	**pcpM** - LysR family transcriptional regulator
	26	**pcpA** - Dioxygenase
ATTCGATGAACGAATTGTGCCTATCCG	21	**pcpE** - maleylacetate reductase
ATCCGACAACTGGAATCTGCTTATCCG	5	**pcpC** - glutathione S-transferase
ATCCAGTATGTGAATGACTTTCATCCA	23	pcpX - aromatic hydrocarbon degradation
ATACTGAATGCGAATGAAGATCATTCG	23	pcpX - aromatic hydrocarbon degradation
ATCCGCGTGACGAATATTCTCTATTCA	21	pcpZ -TonB-dependent receptor plug
	18	phospholipase/carboxylesterase
ATGGTGAATGTGAATGAACCTCATTCG	20	pcpY - TonB-dependent receptor

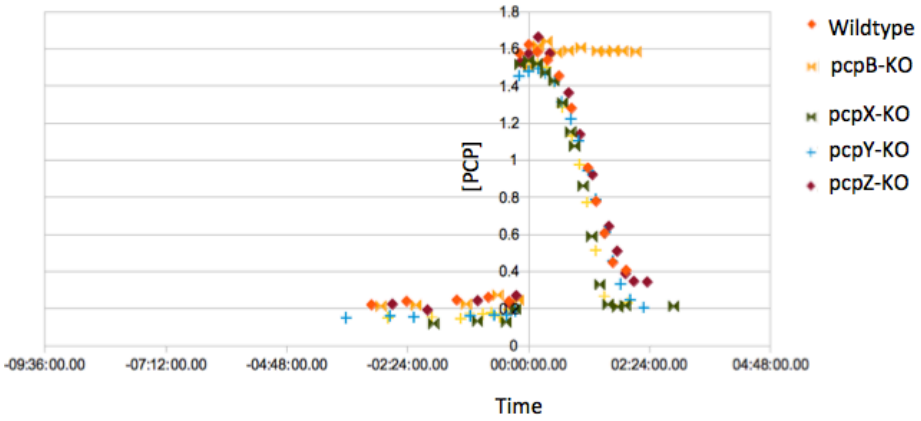
Genes with putative PCP motif in promoter that are also >20 fold induced in the presence of PCP by RNA-seq

The gene *pcpX* actually shows up twice on the above table because it has two full instances of the 3 inverted repeat PCP LysR motif in its promoter region.

Knocking out transporters does not result in a PCP phenotype

The three novel proteins which are over 20 fold induced by PCP and have the PCP LysR motif in their promoters are all annotated as performing some kind of transporter-like function. They are annotated as "TonB dependent receptors" in the case of *pcpY* and *pcpZ* and as an "outer membrane porin" in the case of *pcpX*. I created knockouts of all of these genes and then tested if I could observe a PCP degradation phenotype. I did not find a significant PCP degradation phenotype.

Figure 5.3 - PCP degradation after knocking out putative PCP transporters



Rate of PCP degradation by four knockout strains and one wild type strain of S. chlorophenolicum grown in minimal media. PCP was added to 100 uM at time 00:00. Time in hours and minutes (HH:MM).

It is possible that these transporters perform redundant functions and double or triple knockouts will be required to find a PCP degradation phenotype. Alternatively, these genes may play roles like importing cofactors like iron to replenish cellular stores after this element is depleted by the massive expression of *pcpD*. It is also possible that the role these transporters play in PCP degradation will only be apparent in certain nutrient limiting conditions. Finally, these transporters may play no role in PCP degradation and are relics associated with a PCP gene's ancestral function.

In Vivo Experiments

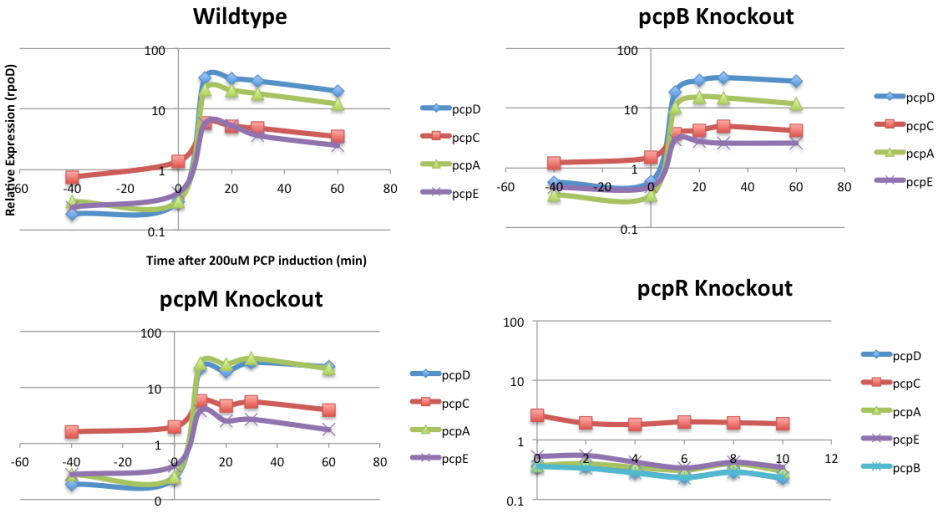
Knockouts of *pcpR* and *pcpM*

In the experiments by Cai et al. the transcriptional effects of *pcpR* or *pcpM* knockout on gene expression were assayed through qualitative RT-PCR products examined on a gel. These assays were not quantitative and were insensitive to all but the most dramatic changes in

expression. I generated *pcpR* and *pcpM* knockout strains and induced them with PCP, taking several time points before and after PCP induction. I also assayed wild type cells to establish base line induction levels and a *pcpB* knockout strain to test if gene induction could occur in the absence of any of the PCP breakdown intermediates (see Figure 5.4). All strains were constructed as described in the Methods.

Induction of the PCP genes in the *pcpB* knockout established that wild type induction of all induced genes occurs in the absence of any PCP breakdown intermediates likely indicating that PCP itself functions as the coinducer of PcpR. I replicate the observation of Cai et al. that PCP induction occurs normally in the *pcpM* knockout, but is completely ablated in the *pcpR* knockout. With the higher sensitivity of qPCR, I observe that induced *pcpA* transcript levels appear to increase in the *pcpM* knockout cell line. Finally, I observe that *pcpC,* previously thought to be constitutively expressed and unregulated, is in fact weakly induced in the presence of PCP, as implicated in the previous section by the presence of the PCP motif. This induction is PcpR dependent, disappearing in the absence of *pcpR*.

Figure 5.4 - PCP induction in knockout strains

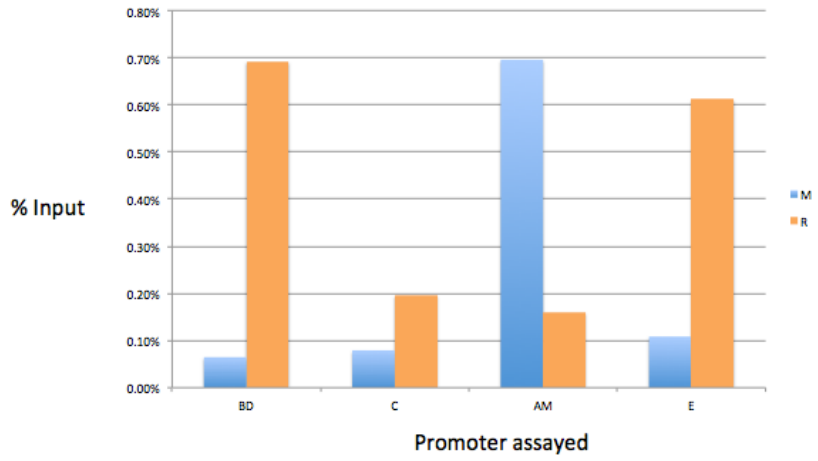


Fold change of pcpD, pcpC, pcpA and pcpE relative to rpoD expression in three different knockout cell lines and wildtype versus time in minutes. Cells induced with 200uM PCP at time 0 min.

ChIP-qPCR for PcpR and PcpM

Given the conservation of PcpM and its apparent regulatory effect on *pcpA*, I undertook an experiment to measure the interaction of PcpM with the PCP degradation promoters in-vivo. I also assayed PcpR interaction with the PCP degradation promoters. Both measurements were performed by constructing strains with endogenously his-tagged versions of PcpM or PcpR, as described in the Methods. ChIP-qPCR with an anti-HIS antibody as described in the Methods was used to assay for protein binding at the promoters of interest in vivo (see Figure 5.5).

Figure 5.5 - ChIP qPCR for PcpR and PcpM at the PCP induced genes



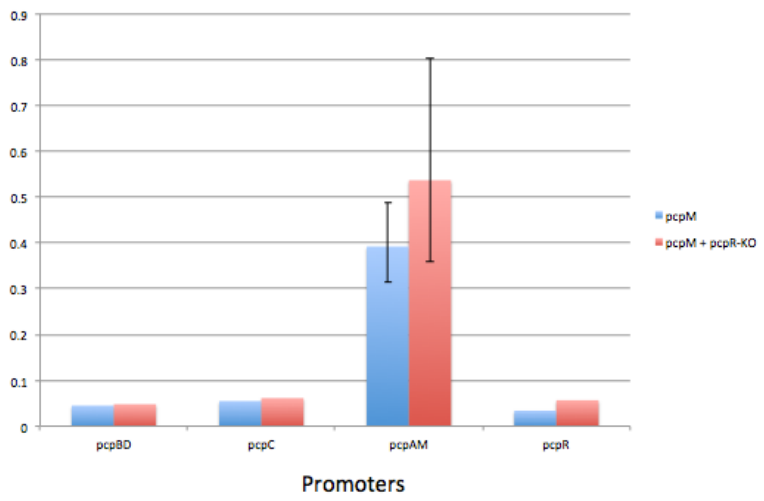
ChIP qPCR representing the occupancy of PcpM (blue bars) and PcpR (orange bars) at the pcpBD, pcpC, pcpAM and pcpE promoters.

I was surprised to discover that PcpR binds strongly at the *pcpBD* and *pcpE* promoters but fails to bind strongly at the *pcpAM promoter.* PcpM shows an inverted binding pattern. It binds strongly at the *pcpAM* promoter but fails to bind significantly at the *pcpBD, pcpC* or *pcpE* promoters. This is surprising for several reasons. First, this data seems to suggest that PcpM is outcompeting PcpR at the *pcpAM* promoter in wild type cells despite observation that PcpR both necessary and sufficient for *pcpA* induction and the observation that pcpM is neither necessary nor sufficient to induce *pcpA* expression in the knockout experiments.

PcpR does not prevent PcpM from binding

If PcpM is outcompeting PcpR at the *pcpA/M* locus, it is plausible that the reverse is occurring at the *pcpBD* and *pcpE* loci. To test this hypothesis, I constructed a his-tagged PcpM strain in which *pcpR* was also knocked out (Figure 5.6).).

Figure 5.6 - ChIP qPCR for PcpM in wild type and pcpR knockout backgrounds



%Input is plotted indicating ChIP enrichment of his-tagged PcpM at the pcpBD, pcpC, pcpAM and pcpR promoters relative to input. Blue bars are his-tagged PcpM in a wildtype strain. Red bars are his-tagged PcpM in a pcpR knockout strain.

Knocking out *pcpR* does not rescue *pcpM* binding at the other loci. Even in the absence of competition from *PcpR, PcpM* is unable to bind at any promoters besides its own. There must be some mechanism by which PcpM is able to bind to its own promoter to such a degree that it can outcompete PcpR, yet at the same time fail to outcompete PcpR for binding a few kilobases away at the *pcpE* promoter. The mechanism is likely something unique about the sequence of the pcpA promoter itself or a local effect resulting from the fact that pcpM is transcribed and translated at this location. I have shown that pcpM binds its own promoter in vivo. There are several lines of reasoning that suggest that pcpM binds to the PCP motif. First, there is no other LysR promoter in the intergenic region for pcpM to bind. Crystal structures of LysR DNA binding domains show which residues of in the protein sequence specifically contact the DNA (Alanazi, Neidle, and Momany 2013). These residues are identical in PcpM and PcpR. Finally, *pcpM* appears to be preventing *pcpR* from binding the DNA indicating competition for

the same binding site. To measure this, I undertook in vitro experiments with purified protein and DNA.

In Vitro Experiments

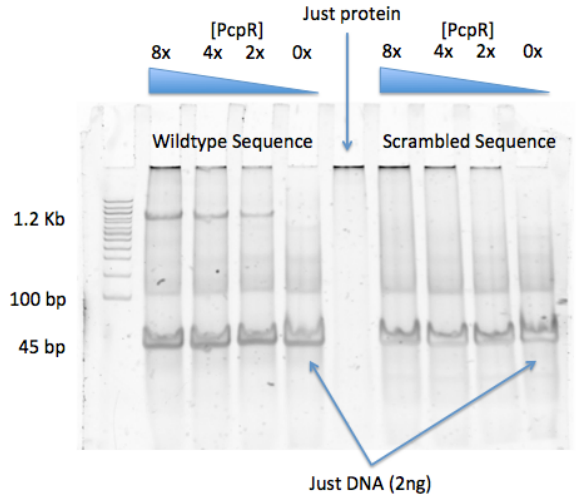
PcpR binds the PCP motif

Before attempting to quantify specific binding constants with more labor intensive radio labeled gel shifts, I ascertained the relative binding characteristics of PcpR to a broad range of synthesized DNA molecules by performing EMSAs in small format acrylamide gels stained for DNA with Sybr-SAFE and then stained again afterwards for protein with coomassie.

I assumed a dissociation constant in the 500 pM range. Visualizing DNA with the Sybr-SAFE stain required loading DNA in the high nM range many hundreds of time higher than its expected dissociation constant. Unfortunately, this necessitated loading purified protein in the uM range at which concentration aggregates would form easily. This prevented the measuring of binding constants but provided a low cost and rapid way to examine many different binding conditions with many different DNA substrates.

To establish that I could measure protein DNA interactions with this assay, I took freshly purified protein at ~1uM and allowed it to bind to synthetic oligos encoding ~45bp sequences centered on the three inverted repeats of the *pcpB* promoter's LysR motif. A negative control in which the nucleotides within the three inverted repeats were scrambled was also tested to establish sequence specificity while controlling for GC content and overall nucleotide composition (see Figure 5.7).

Figure 5.7 - PcpR binds the pcpBD promoter sequence specifically

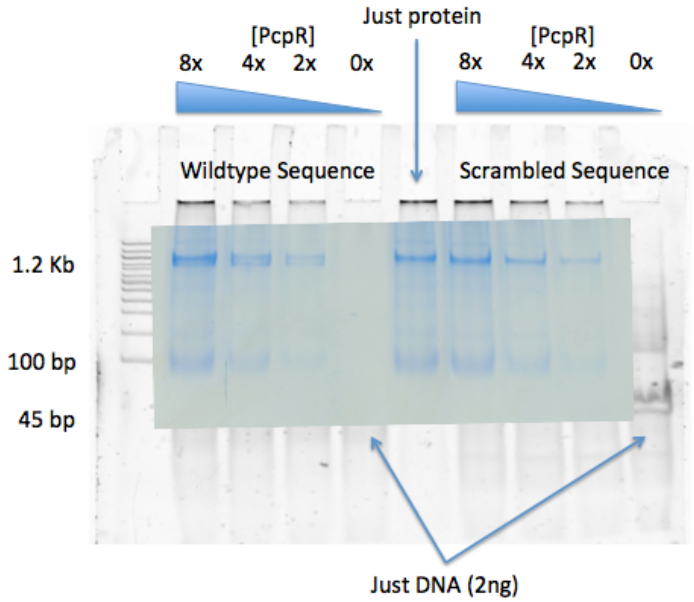


Sybr-SAFE EMSA. Lane 1 is a 100bp ladder. Lanes 2-4 contain a 2 fold serial dilution of PcpR protein starting with ~1uM PCP. Lane 5 contains no protein and only DNA. Lanes 2-4 contain 45bp oligo centered on the PCP motif in the pcpB promoter. DNA is added at a concentration of 25uM. Lanes 7-10 are analogous to Lanes 2-5 but the 12 base pairs that make up the three inverted repeats have been scrambled.

In Figure 5.7, the large band at the bottom near 45 bp represents unbound DNA oligo. The band at 1.2 kb represents DNA that has bound the protein. This shifted band, which signifies the protein-DNA interaction, disappears completely in the absence of PcpR protein in the "0x" lane. Additionally, this shifting is shown to be sequence specific to the putative LysR motif because it also disappears when the sequence is scrambled. This is the first time the PcpR protein has been shown to interact with the putative motif.

In Figure 5.8, the gel is stained for protein with coomassie and superimposed onto the Sybr-SAFE stained gel.

Figure 5.8 - PcpR EMSA stained with coomassie



The gel in Figure 5.7 stained with coomassie and super imposed over the Sybr-SAFE stained gel.

One strange phenomenon is immediately apparent. Regardless of the presence of DNA, the protein at the shifted band migrates to almost the exact same position. This suggest that the location of band signifying the protein DNA interaction is dominated by the mobility of the protein in the native gel rather then being determined by the negative charges of the complexed DNA. Supporting this conclusion, protein migrates to only one place in lanes mixed with scrambled DNA but protein mixed with unscrambled DNA yields double band. This double band is barely visible at the resolution of the gel and is not present in the DNA stained gel. I interpret the double band as two populations of PcpR protein. One population consists of unbound protein migrating natively to this position in the gel. The lower band consists of bound protein pulled deeper into the gel by the negative charge of the DNA. At equilibrium with these concentrations of protein and DNA, neither all of the DNA, nor all of the protein, is bound.

Interactions with the pcpB promoter after scrambling individual repeats.

Many LysR type regulators bind DNA as tetramers. The DNA binding motif consists of one pair of inverted repeats that function as a recruitment binding site (RBS). Nearby are another pair of inverted repeats with many substitutions relative to the consensus sequence. This second more degraded pair acts as the activation binding site (ABS). As the protein only contacts the DNA via the four identical DNA binding domains present in the tetramer, it would stand to reason that disrupting one inverted repeat would have the same consequence as disrupting any other inverted repeat. This turns out not to be the case though.

I tested if all or only a subset of the three inverted repeats were necessary for PcpR binding in this assay. If we number the repeats 1 through 3 from left to right, I scrambled only repeat number 2 in one oligo and only repeat number three in another oligo. When the experiment was repeated it was seen that scrambling the middle repeat almost completely ablated the DNA binding interaction. In contrast, scrambling the rightmost inverted repeat had no apparent affect on the binding interaction.

This result is consistent with a model in which PcpR binds as a dimer and tetramerizes on the DNA. Alternatively, it is possible that two adjacent repeats must be bound by the protein complex to make the third site accessible either through a conformational change induced in the complex or one induced in the DNA.

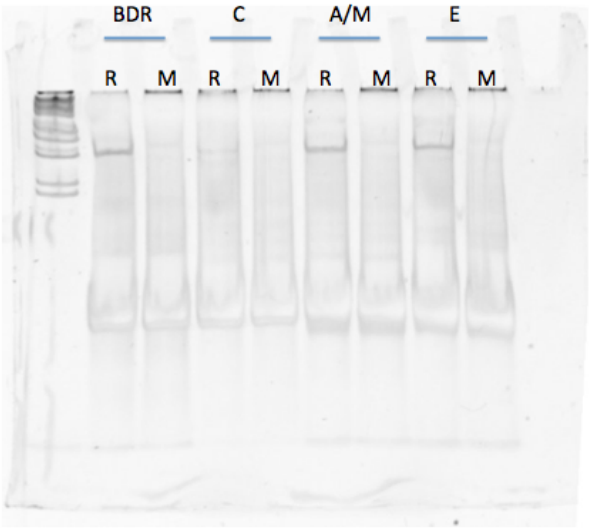
Interactions with the pcpB, pcpC, pcpM/A, and pcpE promoter motifs

Finally, it has been shown in other LysR regulators that DNA residues outside of those that interact with the sequence recognition coil of the DNA binding domain of the protein can have an effect on the ability of a regulator to induce transcription. In ArsR for example, a poly A

tract between inverted repeats is necessary for full induction of the regulated gene (Porrúa et al. 2013). Because all of the PCP induced genes are induced to different levels, it might be possible to explain this behavior through differential binding of PcpR to the different specific sequences, both through changes in the core motif and in the sequences between the inverted repeats.

To test this possibility, I repeated the experiment using oligos centered on all of the putative motifs. With the exception of the *pcpC* promoter motif, the sequences appear to have the same amount of binding. The *pcpC* promoter oligo is significantly less effective at binding the PcpR protein. No interaction is apparent between any of these DNA sequences and the PcpM protein (Figure 5.9).

Figure 5.9 - PcpR binds the motif at the pcpBD, pcpAM and pcpE promoters



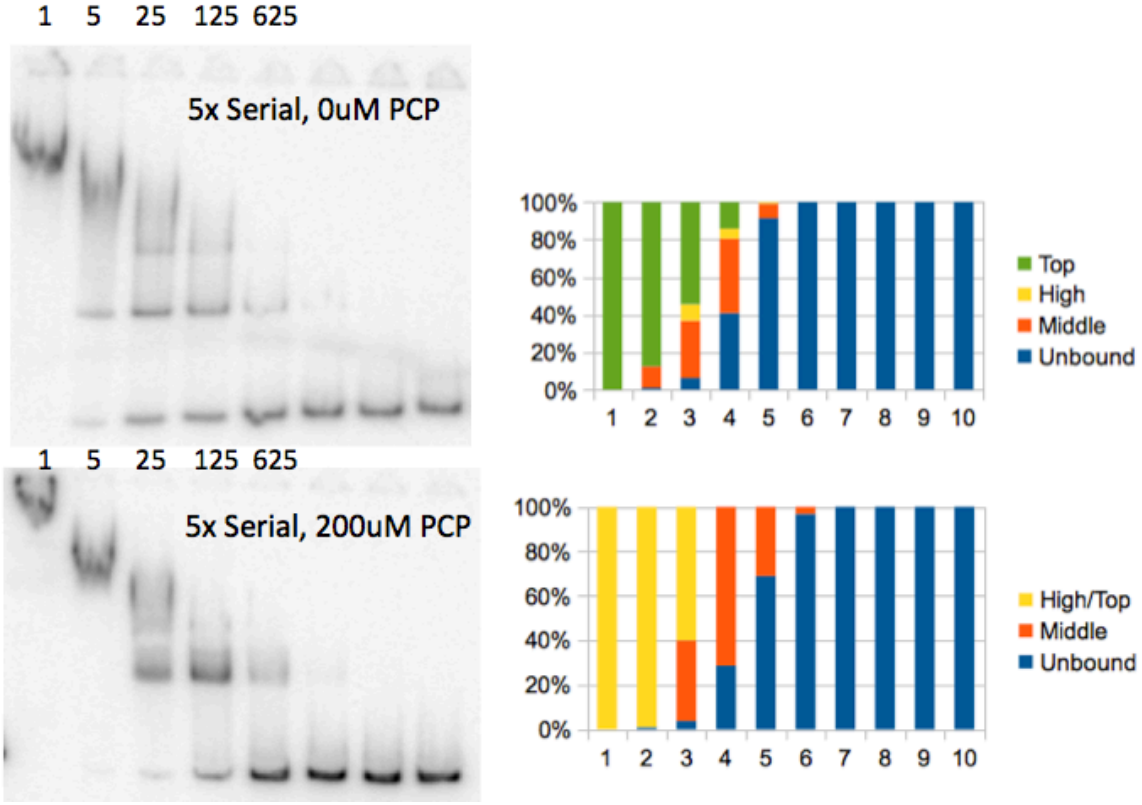
Number from the left: Lane 1 is an 100bp ladder. Lanes 2,4,6,8 contain 1uM PcpR. Lanes 3,5,7,9 contain 1uM PcpM. The DNA bound to the protein are 45 bp synthetically produced oligos centered on the PCP motif in the pcpBD (lanes 2 and 3), pcpC (lanes 4 and 5), pcpAM (lanes 6 and 7) and pcpE (lanes 8 and 9) promoters.

This pattern is consistent with the observation that the core motif of the *pcpC* promoter deviates significantly from the consensus motif. Also, in vivo, *pcpC* is constitutively expressed and after being induced, it is induced to a much lower level than the rest of the induced PCP genes. Perhaps this lower final induction level is caused in part by a lower affinity of PcpR for the *pcpC* promoter. Alternatively, the MEME recovered motif could be different from the one responsible for PcpR dependent induction. There is a separate imperfect motif identified by Cai et al. in these regions at about the -55 region. This motif is GTTC-N_7-GAAT. It is possible that this motif is responsible for induction at *pcpC* by PcpR. Further experiments are required to distinguish these possibilities.

Radiolabeled EMSA of PcpB oligo

After establishing that PcpR and the DNA interact sequence-specifically, I designed equivalent experiments for radiolabeled DNA in order to allow detection of pM concentrations of DNA. These concentration ranges are below the predicted dissociation constant of the protein, allowing us to directly measure the binding constant. I performed a gel shift with 5 fold serial diluted PcpR protein and combined with 100 pM 45 base pair pcpB promoter PCP motif DNA. I established that at these much lower DNA concentrations an interaction still occurred but, rather than a single protein shifted band and a single unbound band, I observed multiple protein shifted bands in addition to the one unbound band (see Figure 5.10 - top panel). By titrating the protein, I was able to go from 100% shifted protein to 100% unbound protein with the DNA transitioning through the various intermediate bands on the way. I believe the top, ill-defined band corresponds to aggregate that retain DNA binding ability. The other two bands likely correspond to a tetramer and dimer population of protein.

Figure 5.10 - PcpR and the pcpB promoter oligo in the presence of PCP



Each lane corresponds to a 5 fold dilution of PcpR protein relative to the lane to its left with the exception of the rightmost lane that contains no protein at all. The top gel was in the absence of PCP, the bottom gel was in the presence of PCP. Bands are quantified and plotted as a bar graph showing the present of signal in either the top, high, middle, or bottom band. In the gel shift with PCP top and high bands and quantified together as they could not be resolved.

The cognate inducer of a LysR type regulator may induce a conformational shift in the protein or oligomerization that influences affinity for DNA. I tested if pentachlorophenol affects the binding of PcpR to the *pcpB* promoter by performing the same experiment described above in the presence of 200 uM pentachlorophenol.

As can be seen in Figure 5.10, there are many qualitative differences in the presence of PCP. Instead of forming distinct top, high, and middle bands, in the presence of PCP the bands between top and middle form a more amorphous blur.

The bands were quantified in ImageJ. While there is a distinct unbound band both in the presence and absence of PCP, the high and top bands run together in the presence of PCP. Combining these categories and instead just looking at the transition from completely bound to unbound it appears that the kd is very similar with or without PCP. The defining distinction therefore appears to be that the middle and high bands run differently in the presence of PCP.

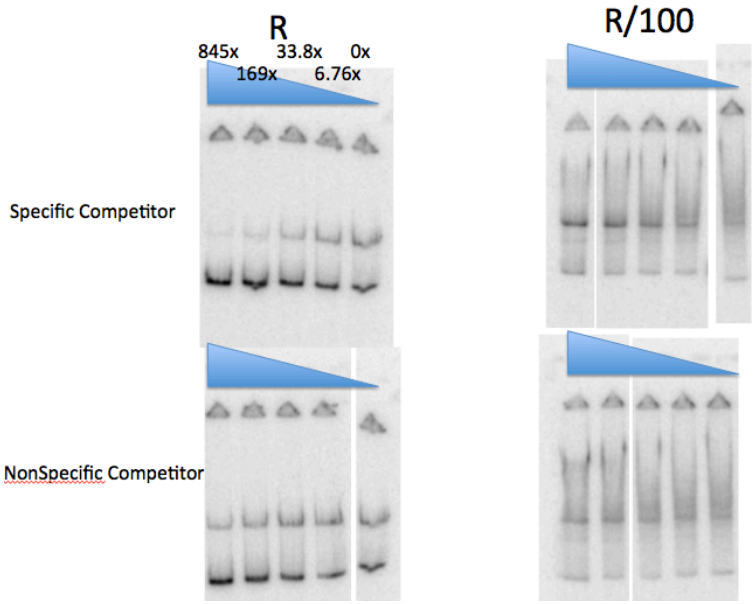
Gel shifts with unlabeled competitors

After establishing that I could recapitulate a DNA protein interaction in the radiolabeled gel shift, I attempted to show that this interaction was sequence specific by competing it off with unlabeled specific and unlabeled nonspecific DNA competitors. I performed the competition assay at two concentrations of PcpR and titrated the levels of competitor. Protein was allowed to reach equilibrium with the competitor before the labeled probe was added and equilibrium was reestablished before the binding reactions were loaded onto the gel.

The higher of the two PcpR concentrations tested behaved as expected. Increasing amounts of specific but not nonspecific competitor displaced the labeled probe. Strangely, at the lower concentration of PcpR the opposite effect was observed. Increasing amounts of specific competitor increased the amount of shifted band. Increasing amounts of nonspecific competitor had no effect (see Figure 5.11).

I interpreted this phenomenon as resulting from a reversible aggregation of the PcpR protein. Specific competitor disrupts this aggregation, increasing the concentration of protein released from the aggregate. As a result, there is a higher concentration of free protein in the presence of more specific competitor and so I see more binding, as opposed to less, and specific competitor is increased.

Figure 5.11 - Specific and Nonspecific DNA competitors at two concentrations of PcpR



Blue triangle represents increasing amounts of specific competitor with ~1 uM on the left, 5 fold serial dilution for the next 3 lanes and none on the left. The gels on the right contain 100 times less PcpR protein than the gels of the left. Protein concentration is fixed in any given gel.

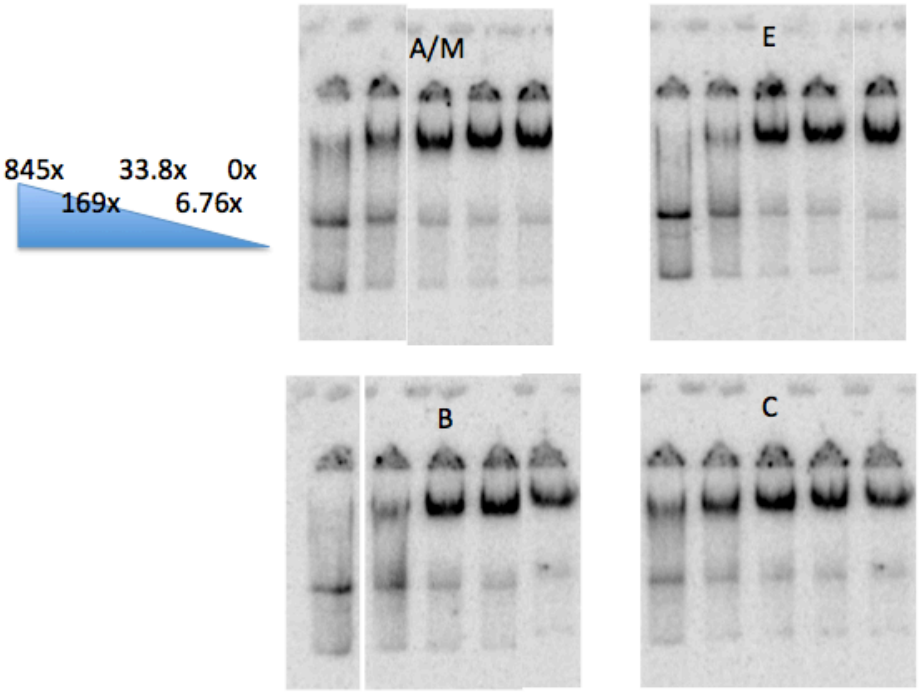
Higher order oligomerization of PcpR and PcpM

Full length LysR regulators are known to precipitate from solution at high concentrations (Ezezika et al. 2007). This fact has hindered the study of this largest known class of transcriptional regulators as it precludes generating crystal structures or performing gel shifts which both require concentrations of protein much higher than are present in vivo. This precipitation can be mitigated, at least partially, through truncation of the DNA binding domain. By truncating this piece of the protein and leaving behind only the effector binding domain, precipitation can be avoided. Recently, it was proposed that many LysR regulators oligomerize by first forming a dimer and then a dimer of dimers to form a tetramer. Tetramerization however, leaves two unsatisfied tetramerization interfaces. At low concentrations, this is not

an issue, but at higher concentrations, the LysR regulators can form into large unbounded linear arrays of protein (Ezezika et al. 2007). These linear arrays can daisy chain together in one of two ways depending on how the DNA binding domains of adjacent monomers interact. In one form, tetramers are connected together and the DNA binding domains do not interact across these tetramerization boundaries. In the other form, the DNA binding domains bridge the tetramerization interfaces and result in a insoluble complex that can not easily dissociate into its component tetramers.

It has been noted that adding a DNA scaffold for the protein to bind allows the protein to remain soluble at much higher concentrations (Ruangprasert et al. 2010). During gel shifts to obtain binding constants of the PcpR and PcpM proteins to various DNA molecules, I noted this effect. In the following experiment, I added the same amount of PcpR protein to a variable amount of four different unlabeled competitor oligos. After the protein was allowed to reach equilibrium, I added in 100pM labeled probe (Figure 5.12).

Figure 5.12 - PcpR binding to pcpB motif oligo in presence of competitor



PcpR protein mixed with a 5x serial dilution of one of four different unlabeled specific competitors. The most highly concentrated competitor is the lane on the left at ~8.5uM. The rightmost lane of each gel contained no competitor. After equilibrium the samples are mixed with 100 pM labeled pcpAM probe and run on an acrylamide gel. The blue triangle on the left indicates the ratio of unlabeled competitor to labeled probe. Starting in the upper left and moving clockwise, the gels contain unlabeled competitor oligos of centered on the PCP motif from the promoters of pcpAM, pcpE, pcpC and finally pcpB.

In some cases, adding competitor DNA actually increased binding of the protein to the labeled probe. This phenomenon can only be explained if the presence of competitor DNA shifts the equilibrium of protein from the higher ordered linear network into the more soluble dimers and tetramers increasing the concentration of these species.

Conclusion

In this chapter, I showed that PcpR interacts sequence-specifically with its putative motif. I provided evidence that *pcpM* is transcribed and translated and binds to its own promoter. I showed that PCP does not significantly affect the dissociation constant of PcpR to

DNA in vivo. I provided evidence suggesting that PcpM is blocking PcpR from binding, but that the converse is not true for PcpM.

Unfortunately, I was unable to measure dissociation constants for these proteins due to difficulties with the proteins aggregating. However, I did show that, in the case of PcpR, DNA containing the PCP binding motif likely inhibits this aggregation.

Chapter 6 Summary and Conclusion

In this thesis, I performed genome sequencing, RNA sequencing, comparative genomics, and targeted biochemical experiments to better characterize the PCP degradation pathway of *S. chlorophenolicum* with the hope of providing evolutionary and genomic context to better understand the evolutionary trajectory that led to this pathway. Each approach yielded significant new discoveries.

By sequencing the genome, I found that, in spite of expectations, there were no close paralogs of the PCP genes encoded in the genome. The presence of close paralogs would have been indicative of duplication and divergence. The absence of paralogs instead suggests recruitment of these enzymes as a more likely explanation. Although, I did not find any close paralogs in the genome, I did find an intriguing pattern of conservation between homologous PcpE and PcpA proteins in the *S. chlorophenolicum* and *S. japonicum* genomes suggestive of an ancient duplication and divergence event. I found that *pcpB*, *pcpD* and *pcpR* have anomalous GC content, a possible sign of HGT. However, in contradiction, I also found putative orthologs of all of the PCP genes in the recently sequenced *Novosphingobium resinovorum* genome, suggesting a pattern of vertical descent. More genome sequences are needed to unambiguously determine the evolutionary history of these genes.

Transcriptomics allowed us to define many new sets of genes that respond to PCP stress. In addition to the known PCP degradation genes, I observed the induction of a diverse set of genes involved in functions ranging from transport and efflux to the degradation of related aromatic compounds.

Our investigations into the molecular biology of PcpR and PcpM also yielded new information. I showed that PcpR directly and sequence-specifically interacts with the ATTC-N_7-GAAT motif present in the *pcpA*, *pcpE, pcpC* and *pcpBD* promoters and that PCP does not appear to affect the affinity of PcpR for DNA. I showed evidence that *pcpM* is not only transcribed but also translated into a functional protein capable of binding its own promoter in-vivo. I provided evidence suggesting that PcpM blocks PcpR from binding at the promoter of *pcpA*, but that the converse does not appear to be true: even in the absence of PcpR, PcpM does not bind any of the other PCP gene promoters.

I found that one the *pcpE* homologs uncovered during the genomic analysis is part of a three gene operon that is upregulated in response to PCP. This operon includes a monooxygenase and a hypothetical enzyme and is transcribed divergently from a LysR type regulator. The intergenic region upstream of this operon does not contain a PCP regulatory motif, suggesting that this operon is likely induced by some other mechanism perhaps promiscuous activation of the adjacent LysR regulator.

By examining the genomic DNA for overrepresented motifs, I was able to expand the known PCP motif. I exhaustively scanned the genome for genes with this new motif near their promoters. I then cross-referenced these genes with the RNA-seq data set to compile a list of gene candidates likely induced by PcpR.

Through qPCR of wild type and *pcpR* knockout strains, as well as ChIP-qPCR, I showed that the *pcpX* gene is induced in a PcpR dependent manner and that PcpR binds at its promoter in-vivo. Finally, through this same approach showed that PcpR regulates *pcpC* as well, a gene previously thought to be constitutively expressed and unregulated.

We do not have a genome sequence for the ancestor of *S. chlorophenolicum*, nor do we know the ancestral role the PCP degrading enzymes played before the global selection pressure of industrially produced PCP. But these results fill large gaps in our understanding of this pathway and how it connects to many new or previously known genes.

Much remains to be studied. The *pcpX, pcpY* and *pcpZ* transport related proteins are all tantalizing new subjects to investigate. Our preliminary experiments show that individually disrupting these genes does not affect PCP degradation but do these proteins perhaps overlapping roles in the degradation of PCP? Are they relics of the ancestral functions of the PCP genes?

The role of PcpM in regulating the PCP degradation pathway is also still a mystery. I found that PcpM binds its own promoter and likely at least partially prevents PcpR from binding there, but the functionality of this behavior remains unclear. Perhaps most baffling of all, is how PcpM manages to bind its own promoter but does not bind to other PCP gene promoters even when the same motif is present in all of them. Is this a consequence of the slow decay of PcpM as it destabilizes into a pseudogene in the absence of selective pressure, or is this a mechanism that allows it and PcpR to share a motif without interfering with each others' functions? The presence of orthologs of these regulators in the distantly related Novosphingobium suggests that whatever these two regulators are doing together they have been doing it for a long time in both lineages.

The pursuit of understanding the process of microbial evolution is an essential one. Diverse in it's application, this process simultaneously bewilders clinicians as antibiotic resistance becomes frequent and inspires synthetic biologists as directed evolution is applied to

engineer new enzymatic pathways. By studying the PCP degradation pathway of *S. chlorophenolicum,* hopefully we can learn lessons than can be applied to further our understanding of this powerful process.

Chapter 7 Methods

S. chlorophenolicum Genome Modifications

E. coli plasmid origins of replication do not replicate in *S. chlorophenolicum*. This property allows for an extremely simple method for generating genomic mutations. A *S. chlorophenolicum* knockout is created by electroporating electrically competent *S. chlorophenolicum* with ug amounts of plasmid DNA purified from *E. coli.* The plasmid DNA contains a region of homology upstream of the target area to knock out, a kanamycin or hygromycin resistence cassette, and finally a region of homology from downstream of the target area to knock out. The plasmid is unable to replicate in *S. chlorophenolicum* and in order to incorporate the resistance marker a double recombination event must occur.

Generating the Knockout Plasmid

Four regions were combined via Gibson assembly to back a knockout. The first is a 500 bp region encompassing the origin of replication from PUC19. This region is flanked by the sequences “TTTCCATAGGCTCCGCCCCCCTGACGAGCA” and “TTGAGATCCTTTTTTTCTGCGCGTAATCTG”. After the origin of replication region is a 500 bp homology region from upstream of the region to be knocked out. Then a kanamycin resistance marker cloned from pACYC184 flanked by the sequences

“ATCTGATCCTTCAACTCAGCAAAAGTTCGA” and “TTAGAAAAACTCATCGAGCATCAAATGAAA” This region includes a kanamycin resistance gene and its associated constitutive promoter. Finally, another 500 bp homology region from downstream of the segment to be knocked out is included. The final plasmid is composed of these four segments: Left homology region, KAN cassette, right homology region, ORI.

The cloning strategy I used to combine these segments into a plasmid was the one step Gibson Assembly (Gibson et al. 2009) following manufacturer’s instructions. The manufacturer’s protocol is described in brief in the following section.

Primers for the four segments of interest were designed using the NEBbuilder program website. This website automatically generates primers with 3’ overhangs such that the final PCR products have >20bp overlapping identical regions at every end that is to fuse together. Genomic or plasmid templates were PCR amplified using these primers. The PCR yield was quantified with a high sensitivity DNA Qubit assay and converted to molar amounts. While the manufacturer’s instructions recommend a molar excess of insert to backbone, because all of our fragments are roughly the same size, I opted for 1:1 molar ratios for all fragments. If PCR fragments are diluted such that they make up less than 20% of the final Gibson reaction volume then it is unnecessary to perform a PCR cleanup before proceeding to the Gibson reaction. For every knockout attempted, the PCR reaction was efficient enough to avoid the necessity for PCR purification. The 20% of the 20uL Gibson reaction is 4 uL. The minimum pmols added was .2 pmols of each fragment. If the sum of the volumes of 0.2 pmol of each fragment was less than 4uL then no PCR purification was necessary. The PCR products were combined in appropriate proportions and then 4 uL were added to 6uL of H20 and 10uL of Gibson reaction

mix. This mix was incubated at 50 degrees C for 15 minutes and then transformed into *E. coli* using commercial high efficiency NEB chemically competent cells. After recovery in SOB for 1 hour, 100 uL of cells were plated on 50 ug/mL Kan LB plates.

Typically, about 100 colonies would grow up. Five transformants were screened by inoculating the colonies in 5 mL test tubes of LB. 10 uL of this was boiled in 1 mL of water and used as template for diagnostic PCR reactions with the forward primer of the left region and the reverse primer of the right. By this check, typically ~80% of transformants resulted in the correct product. Alternatively, in some cases, diagnostic restriction digests were used in place of diagnostic PCR.

Next, large amounts of plasmid must be prepared. Because the plasmid will not replicate inside of *S. chlorophenolicum*, microgram amounts must be added to the electroporation reaction to produce sufficient transformation and recombination efficiency. For such large amounts of plasmid DNA, a miniprep is not sufficient. The most common methods for the purification of large amounts of plasmid are CsCl gradients and ionic exchange columns like the Qiagen midi or maxi prep kits. However, I found that the Qiagen Plasmid midi-plus kit provided large amounts of sufficiently pure plasmid DNA much more quickly. Unlike the Midi and Maxi kits, the plasmid-plus kit is a modification of the silicon dioxide and chaotropic salt method used in common minipreps. Detergents are used to expand the method to about 10 times the number of cells as is feasible with a normal miniprep.

25 mL of LB were inoculated with the PCR confirmed mutant. This was processed using the manufacturer's instructions for the Qiagen Plasmid-Plus Midi Prep. This DNA was eluted in

100 uL of water and concentrated in a SpeedVac until it was at a concentration of at least 1ug/uL.

Transforming *S. chlorophenolicum*

To transform cells, a 25 mL culture of *S. chlorophenolicum* was grown up in quarter strength trypic soy broth (TSB). At an optical density at 600 nm of 1.0, the cells were washed 3 three times with ice cold ddH20 spinning for 10 minutes at 3600 RCF. The final pellet was brought to a volume of 125uL with ice cold ddH20. Ideally the OD of the cells should be 200 although in practice the OD was usually closer to 170 due to losses during the washes. These cells are electrically competent.

45 uL of electrically competent cells were combined with 5 uL of 1 ug / uL knockout plasmid DNA. 2mm electroporation cuvettes were chilled on ice. The cells were then electroporated with one pulse of the Ec2 setting. 950 uL of TSB was immediately added and the cells were allowed to recover for 2 hours at 30C in a shaking incubator before 100 uL were plated on TSB plates containing 5 ug/mL Kan. After about 5 days, transformant colonies appear

Finally, to confirm that the genomic mutation has occurred, colonies were picked into 5 mL TSB overnights and grown for 2 days. 1 mL was used for genomic DNA purification using the Promega Wizard kit. This genomic DNA template was PCR amplified using primers outside of the 500 bp flanking regions utilized to generate the mutant. Wildtype genomic DNA was also amplified and a size difference was used to confirm the mutation. If the size difference was not significant enough to detect on a gel then the mutant PCR product was differentiated from the wild type by restriction digest.

S. chlorophenolicum Genome Sequencing

Isolation of genomic DNA from *S. chlorophenolicum L-1*

A sample of *S. chlorophenolicum* L-1 was originally deposited to the American type Culture Collection and designated ATCC 39723. The initial characterization of the enzymes in the PCP degradation pathway was done with the ATCC 39723 strain. However, the ATCC 39723 strain lost the ability to degrade PCP, so a second sample of *S. chlorophenolicum* L-1 was deposited and designated ATCC 53874. Genomic DNA was isolated from this strain.

A 5 mL culture of ATCC medium 1687 (*Flavobacterium* medium) containing 5 µM PCP was inoculated with a colony from a fresh agar plate of half-strength tryptone soy broth containing 50 µM PCP and incubated with shaking at 30 °C overnight. One mL aliquots of cells were harvested by centrifugation at 16,100 x *g* for 1 min. Genomic DNA was isolated using the Aquapure Genomic DNA kit (Bio-Rad, Hercules, CA). Cell pellets from 1 mL of culture were re-suspended in 300 µL Genomic DNA Lysis Solution. The solution was incubated at 80 °C for 5 min to lyse the cells. A solution of RNase A (1.5 µL, 4 mg/mL) was added and tube was inverted 25 times before incubation at 37 °C for 45 min. The temperature of the sample was adjusted to room temperature and 100 µL of Protein Precipitation Solution was added. The sample was vortexed for 20 seconds and then subjected to centrifugation at 13,000 x g for 3 min to remove cellular debris. The supernatant containing DNA was transferred to a clean 1.5 mL tube, incubated on ice for 5 min and then subjected to centrifugation again. The supernatant was then transferred to a clean 1.5 mL tube containing 300 µL 100% isopropanol to precipitate the DNA. The tube was inverted 50 times. The DNA was pelleted by centrifugation at 13,000 x *g* for 1 min. The supernatant was poured off and the pellet was washed with 300 µL 70% ethanol.

The ethanol was poured off and residual ethanol removed by draining the tube onto filter paper for 10-15 min. The pellet was dissolved in 100 µL 10 mM Tris-HCl, pH 7.5, at 65 °C for 5 min and then at room temperature overnight. The quality of the DNA was checked on a 0.8% agarose 1X TAE gel alongside a 1 kb DNA Extension Ladder (Invitrogen, Carlsbad, CA), and the concentration was measured on a NanoDrop spectrophotometer.

Genome sequencing

The draft genome of *S. chlorophenolicum* L-1 was generated at the DOE Joint genome Institute (JGI) using a combination of Illumina (Bennett 2004) and 454 technologies (Margulies, et al. 2005). Three libraries were constructed and sequenced: an Illumina GAII shotgun library that generated 40,283,193 reads totaling 3061 Mb; a 454 Titanium standard library that generated 302,660 reads; and a paired end 454 library with an average insert size of 11.26 +/- 2.81 kb that generated 174,361 reads. In total, 176.8 Mb of 454 data were obtained. General aspects of library construction and sequencing performed at the JGI can be found at http://www.jgi.doe.gov/. The initial draft assembly contained 79 contigs in one scaffold. The 454 Titanium standard data and the 454 paired end data were assembled with Newbler, version 2.3. The Newbler consensus sequences were computationally shredded into 2 kb overlapping fake reads (shreds). Illumina sequencing data was assembled with VELVET, version 0.7.63 (Zerbino, 2008), and the consensus sequences were computationally shredded into 1.5 kb shreds. The 454 Newbler consensus shreds, the Illumina VELVET consensus shreds and the read pairs in the 454 paired end library were integrated using parallel phrap, version SPS - 4.24 (High Performance Software, LLC). The software Consed (Ewing and Green 1998; Ewing et al. 1998; Gordon et al. 1998) was used in the finishing process. Illumina data was used to correct

potential base errors and increase consensus quality using the software Polisher developed at JGI (Alla Lapidus, unpublished). Possible misassemblies were corrected using gapResolution (Cliff Han, unpublished) or Dupfinisher (Han, 2006), or by sequencing cloned bridging PCR fragments. Gaps between contigs were closed by editing in Consed, by PCR and by Bubble PCR (J-F Cheng, unpublished) primer walks. A total of 87 additional finishing reactions (either sequencing of bridging PCR fragments or primer walking) were necessary to close gaps and to raise the quality of the finished sequence. The total size of the genome is 4,573,221 bp. The final assembly is based on 176.8 Mb of 454 draft data that provides an average 36x coverage of the genome and 3553 Mb of Illumina draft data that provides an average 790x coverage of the genome.

Genes were identified using Prodigal (Hyatt et al. 2010) as part of the Oak Ridge National Laboratory genome annotation pipeline, followed by a round of manual curation using the JGI GenePRIMP pipeline (Pati, et al. 2010). The predicted CDSs were translated and used to search the National Center for Biotechnology Information (NCBI) nonredundant database, UniProt, TIGRFam, Pfam, PRIAM, KEGG, COG, and InterPro databases. These data sources were combined to assert a product description for each predicted protein. Non-coding genes and miscellaneous features were predicted using tRNAscan-SE (Lowe and Eddy 1997), RNAmmer (Lagesen, et al. 2007), Rfam (Griffiths-Jones, et al. 2003), TMHMM (Krogh, et al. 2001), and signalP (Bendtsen, et al. 2004).

Electromobility Shift Assays (EMSAs)

I performed two variants of EMSA, a Sybr-SAFE visualized EMSA that was performed at concentrations of DNA well above the disassociation constant of the reaction and a

radiolabeled EMSA performed at concentrations of DNA below the expected dissociation constant.

Preparing the DNA

DNA probes were generated either through PCR of a genomic DNA template or through hybridization of complementary synthetic oligos. If oligos were prepared through hybridization, they were hybridized in 200mM Tris pH 7.9, 20mM MgCl2, 500mM KCl. The reaction is placed in a thermocycler and incubated at 95C for 3 min. The temperature is allowed to ramp down to 4C at -0.1C per second. The resulting annealed oligos can be stored at 4C.

Preparing the Gel

A 1.5mm 5% Tris-Glycine acrylamide mini gel was cast and pre-run in Tris-Glycine running buffer with 5% Glycerol for 30 min. Two gels could be made by mixing 3mL 10x Tris-Glycine Running buffer, 3mL 50% Glycerol, 5mL 30% Acrylimide and 24mL H20 to a final volumn of 30mL in a 50mL conical. When the gel casting apparatus is set up and the gel is ready to be poured 169uL of 10% APS and 41.25 uL of TEMED are added to the 30mLs in the conical. The reaction is inverted several times and then poured into the casting equipment. If problems with leaking are encountered, vasaline on the bottom of the glass where it touches the foam can be added to create a better seal.

The Binding Reaction and Loading the Gel.

The final concentration of DNA in the binding reactions loaded on the SYBR-safe gels was ~25 uM, the lowest possible concentration where DNA could still be visualized after staining with Sybr-SAFE. 10uL of DNA and 10uL of protein were combined in a binding buffer

consisting of 10mM Tris, 4mM MgCl2, 1mM DTT and 20% Glycerol. Alternatively, for the radiolabeled EMSAs the DNA in the binding reactions was 100 pM. I experimented with many different variations of the binding buffer adding more or less Mg, Zn, DTT, Glycerol etc. Glycerol concentration and whether the glycerol is added to the DNA, the protein or both had an effect on the binding. Adding too much zinc precipitated the DNA. Other than that the various concentrations and additives had no effect on the binding.

The gel was prerun for 20min in Tris-Glycine running buffer. Then, the entire binding reactions were loaded, while the gel was running. The gel was run for 25min.

SYBR-safe EMSA

To visualize the SYBR-safe version of the EMSA 10,000x SYBR-safe is diluted to 1x in 50mL of buffer. The buffer and gel are placed in an empty pipette box and agitated slowly on an orbital shaker for at least 10 minutes. The gel was then visualized on a LAS4000 imager.

Radiolabeled EMSA

In the case of the radiolabeled EMSA, DNA is prepared by PCR product or oligo hybridization as described above. After preparing the DNA it is radiolabeled. 2uL of 1uM Annealed Template is mixed with 2uL of 10x PNK (polynucleotide kinase) Buffer, 2uL of T4 PNK enzyme, 1uL of hot ATP and 13uL of H20. The reaction is incubated for 30min at 37C. While the reaction is running, a Prepacked G-25 Sephadex mini column is prepped. To prep a G25 column, crack off the end and spin it at 4000rpm for 2min in a collection tube. Discard the flow through. Next add the labeled DNA to the G25 column to remove unused hot ATP. Place the prepped column in a fresh tube. The entire 20uL reaction is added directly to the column. This

is then spun for 2 min at 4000 RPM. The green die and hot ATP will remain in the column. The flow through contains the purified labeled dsDNA. The binding reaction and gel separation are performed exactly as described above.

To visualize the radiolabeled EMSA, the gel is placed on filter paper and covered with saran wrap. This is placed on a vacuum-sealed gel drier at 50C for 45min. While the gel is drying, a phosphorimaging screen is blanked by exposing to a light box for at least 15 minutes. The dried gel is then transferred to the blanked phosphorimaging screen and left to develop overnight. The phosphorimaging screen is visualized in a Typhoon scanner the next day.

FPLC Purification of PcpR

The open reading frame encoding PcpR was cloned directly from *S. chlorophenolicum* genomic DNA into a Novagen pET44tr vector. This vector appends a 6xHIS tag on the C-terminus of the open reading frame and places the RNA under the control of a T7 promoter. The plasmid was transformed into BL21 (DE3) protein expression *E. coli* strain.

This strain was grown in a 25mL overnight of LB with appropriate antibiotics then used to inoculate 1.5L of LB with appropriate antibiotics. At an OD600 of 0.2 to 0.3 the cells were induced with a final concentration of 300mM IPTG. Induced culture was left at room temperature overnight and harvested in the morning. Significant differences were not detected by inducing for different times, at different ODs, or at different temperatures.

The culture was spun down in aliquots of 250mL in centrifuge bottles at 6000C for 20min at 4C. The pellets were resuspended in TE, combined, and pelleted again for storage at -20C or for immediate use.

To extract the protein, the cell pellet was resupended in Binding Buffer at a ratio of ~4mL of buffer for 1 g of cell pellet. Binding Buffer is 10% glycerol, 5mM Imidazole, 250mM NaCl and phosphate buffer at pH 8. The cells were passed through the French Press at 10,000 PSI 3 times. 100 uL of the lysate was reserved for quality control. The lysate was placed in a small high speed centrifuge bottle and spun at 20,000 x G to remove cell debris. The supernatent was moved to a fresh tube. The pellet was resuspended in the same volume of TE as the supernatent removed and 100uL was reserved for quality control (insoluble fraction).

Before nickel affinity purification of the lysate, a HisTrap FF Crude 1mL column was equilibrated with 5mL of h20 and then 5mL of binding buffer. Again, Binding Buffer is 10% glycerol, 5mM Imidazole, 250mM NaCl and phosphate buffer at pH 8. The clarified lysate was then applied directly to the column by slowly dripping in with a syringe or peristaltic pump at a rate of about 1mL/min. The column is then washed with 10mL of binding buffer with a peristaltic pump at 1mL/min. 100uL aliquots of the wash were saved for quality control. The sample is then placed on the FPCL. The sample was washed with 5 column volumes of 100% Binding Buffer. Then transitioned to 100% Elution buffer over 25 column volumes. The Elution buffer is identical to the binding buffer except with 600mM Imidazole instead of 5mM imidazole.

The PcpR protein came off at around 250 to 350 mM imidazole. The vast majority of the protein was lost in the insoluble fraction. At concentrations above .125 mg/mL the protein begins to precipitate from solution. Protein was stored at 4C or -20C in 50% glycerol. The same protocol was used to purify PcpM.

ChIP-qPCR

Fix, quench and freeze cells

250mL of ¼ strength TSB is inoculated to an OD600 of 0.05 with *S. chlorophenolicum* overnights. At OD600 of 0.3, 37% formaldehyde is added to a final concentration of 1% and cells are allowed to fix for 15 minutes. After 15 minutes, the reaction is quenched by the addition of 2.5M Glycine to a final concentration of 250mM. The reaction is allowed to quench for 5 minutes. Then the cells are pelleted, washed twice with 10mM Tris Buffer pH 8.0, and frozen at -20.

French press and clarify cell lysate

Cells are thawed on ice then resuspended in 6mL of lysis buffer (50mM HEPES pH 7.5, 140 mM NaCl, 1mM EDTA, 1% TritonX, 0.1% Na-deoxycholate, 1x protease inhibitor cocktail) and passed through the French press two times at 10,000 PSI. The resulting lysate is clarified by spinning at 20,000xg for 20min. The supernatant is moved to a fresh tube.

Chromatin immunoprecipitation of the lysate

50 uL of Dynabeads Protein F, cat #100.04 are washed twice with 1 mL PBS containing 5 mg/mL BSA. Beads are resuspended in 250uL PBS+BSA. 5 uL of 1 ug/uL Qiagen Anti His antibody are added to the beads. Beads are rotated overnight at 4°C.

Beads are washed twice with 1 mL PBS+BSA then resuspended in 500uL of lysate and incubated overnight at 4°C. 100uL of lysate is incubated overnight as well for use as input control in the subsequent qPCR reactions.

Bead washing

Beads are washed in 1mL Lysis buffer + 0.1% SDS incubating for 4 min at room temperature. Beads are washed in 1mL Lysis buffer + 0.1% SDS + 500mM NaCl incubating for 4 min at room temperature. Beads are washed in 1mL Wash buffer (10mM Tris pH 8.0, 250mM LiCl, 0.5% NP40, 0.5% Na-deoxycholate, 1mM EDTA) incubating for 4 min at room temperature. Beads are washed in 1mL TE incubating for 4 min at room temperature.

Elution, reversal of crosslinks and purification.

Beads are resuspended in 250uL TE + 1% SDS. Input control is diluted 1:100 in 250uL TE+1% SDS. Both samples are incubated at 65°C for 6 to 8 hours. The DNA is then purified using a standard Qiagen PCR purification kit and eluted in 250uL of water. This process yields about 0.5 ng/uL chromatin immunoprecipitated DNA.

qPCR

Standard qPCR practices are used. Percent input is calculated by calculating the ratio of the qPCR signal from ChIP DNA to the qPCR signal from the input DNA after correcting for dilutions.

RNAseq Library Preparation

Sample Preparation

S. chlorophenolicum 3mL overnights in ¼ strength tryptic soy broth were used to inoculate 125 mL flasks containing 25 mL of ¼ strength TSB in triplicate. These flasks were inoculated to OD of .05 and shaken at 30C until they reached an OD of 0.3. Samples were taken immediately before and 15 minutes after the addition of PCP at a final concentration of 200 uM.

The samples were 2 mL of culture removed with a serological pipette and combined with 4mL of RNAlater Bacterial Solution in a 15 mL conical. The samples were inverted several times to insure complete mixing and then incubated at room temperature for at least 5 minutes, spun down at 4500 RPM, and the pellet briefly dried then stored at -80 as per manufacturer's instructions.

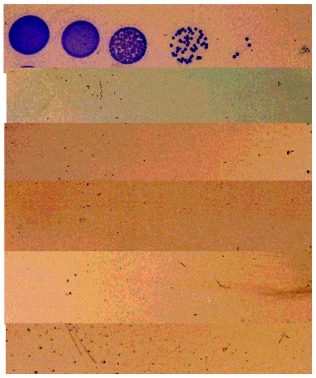
RNA Purification and Library Preparation

Cells were lysed by thawing on ice and incubated in 200uL TE with 5mg/mL lysozyme for 6 minutes (vortexing every two minutes). RNA was purified from the lysate with the Qiagen RNeasy RNA purification kit as per manufacturer's instructions. The total RNA was eluted in 30uL of RNAse free water. The total RNA was checked for quality on a Nanodrop and quantified on a Qubit. Ribosomal RNA was depleted with the Illumina Bacterial Ribo-Zero kit as per the manufacturer's instructions. A multiplexed directional illumina library was generated from the rRNA depleted samples with the NEBUltra Directional mRNA Kit and barcoded using NEBUltra barcoded primers as per the manufacturer's instructions. The library was run on an illumina Hi-seq lane.

Innate Antibiotic Resistance of *S. chlorophenolicum*

Sphingomonads exhibit innate resistance to many classes of antibiotics (Vaz-moreira et al. 2011). Strains show varying levels of resistance to beta-lactams, aminoglycosides, fluroquinolones, polymyxins, sulfonamide. I tested a variety of antibiotics to find which were effective against *S. chlorophenolicum*.

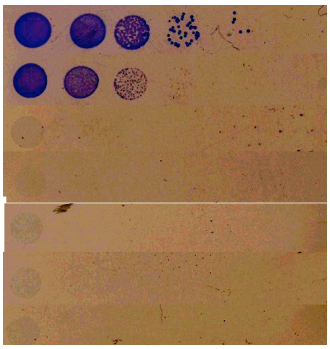
Figure 7.1 - S. chlorophenolicum resistance to kanamycin



Each spot is diluted five fold relative to the spot to its left. The rows from top to bottom correspond to 0, 3.12, 6.25, 12.5, 25, 50 ug/mL kanamycin in ¼ strength TSB agar plates.

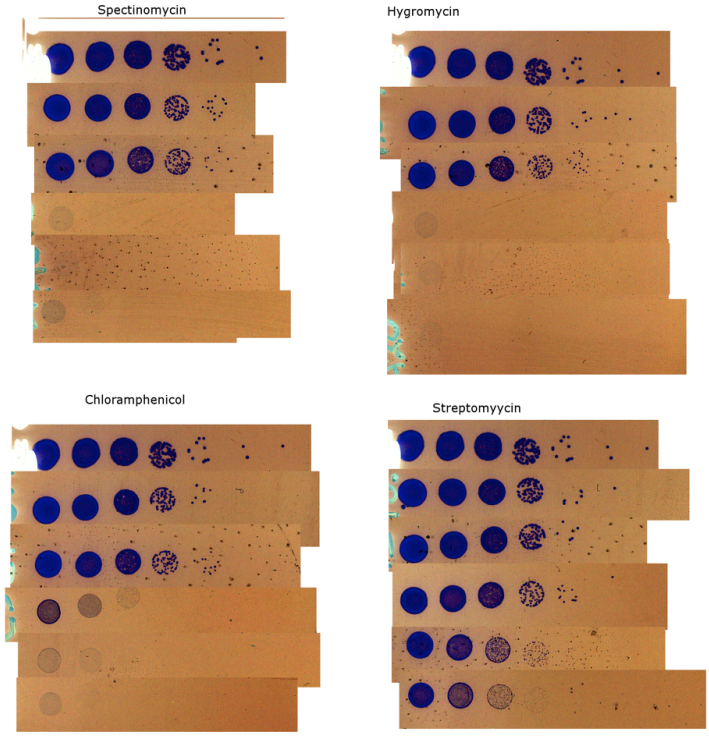
In Figure 7.1, resistance to kanamycin is tested. All growth is stopped even at the lowest concentration of antibiotic (3.12 ug/mL). In contrast, ampicillin is not effective to use against *S. chlorophenolicum*. In the Figure 7.2, substantial growth occurs even at 6.25 ug/mL ampicillin.

Figure 7.2 - S. chlorophenolicum resistance to ampicillin



Each spot is diluted five fold relative to the spot to its left. The rows from top to bottom correspond to 0, 6.25, 12.5, 25, 50, 100, 200 ug/mL ampicillin in ¼ strength TSB agar plates.

Figure 7.3 - S. chlorophenolicum resistance to spectinomycin, hygromycin, chloramphenicol and streptomycin.



Each spot is diluted 5 fold relative to the spot to its left. From top to bottom concentrations are 0, 0.1, 1, 5, 10, 20 ug/mL of the respective antibiotics in ¼ strength TSB agar plates. The antibiotics tested are spectinomycin, hygromycin, chloramphenicol and streptomycin from the upper left moving clockwise.

To make double mutants, it was necessary to find a second antibiotic effective against *S. chlorophenolicum*. I tested spectinomycin, hygromycin, chloramphenicol and streptomycin. Even at the highest concentration of streptomycin, I see growth of *S. chlorophenolicum*. Of the four antibiotics tested here only hyrgromycin and spectinomycin show lethality at similar concentrations to kanamycin.

Bibliography

Aiken, B S, and B E Logan. 1996. “Degradation of Pentachlorophenol by the White Rot Fungus Phanerochaete Chrysosporium Grown in Ammonium Lignosulphonate Media.” *Biodegradation* 7 (3): 175–82. http://www.ncbi.nlm.nih.gov/pubmed/8782389.

Alanazi, Amer M, Ellen L Neidle, and Cory Momany. 2013. “The DNA-Binding Domain of BenM Reveals the Structural Basis for the Recognition of a T-N11-A Sequence Motif by LysR-Type Transcriptional Regulators.” *Acta Crystallographica. Section D, Biological Crystallography* 69 (Pt 10) (October 1): 1995–2007. doi:10.1107/S0907444913017320. http://www.ncbi.nlm.nih.gov/pubmed/24100318.

Altenhoff, Adrian M, and Christophe Dessimoz. 2009. “Phylogenetic and Functional Assessment of Orthologs Inference Projects and Methods.” *PLoS Computational Biology* 5 (1) (January): e1000262. doi:10.1371/journal.pcbi.1000262. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2612752&tool=pmcentrez&rendertype=abstract.

Altschul, S F, W Gish, W Miller, E W Myers, and D J Lipman. 1990. “Basic Local Alignment Search Tool.” *Journal of Molecular Biology* 215 (3) (October 5): 403–10. doi:10.1016/S0022-2836(05)80360-2. http://www.ncbi.nlm.nih.gov/pubmed/2231712.

Anders, Simon, and Wolfgang Huber. 2010. “Differential Expression Analysis for Sequence Count Data.” *Genome Biology* 11 (10) (January): R106. doi:10.1186/gb-2010-11-10-r106. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3218662&tool=pmcentrez&rendertype=abstract.

Apajalahti, JuhaH.A., and MirjaS. Salkinoja-Salonen. 1986. “Degradation of Polychlorinated Phenols by Rhodococcus Chlorophenolicus.” *Applied Microbiology and Biotechnology* 25 (1) (October). doi:10.1007/BF00252514. http://link.springer.com/10.1007/BF00252514.

Baker, Tania a, and Robert T Sauer. 2012. “ClpXP, an ATP-Powered Unfolding and Protein-Degradation Machine.” *Biochimica et Biophysica Acta* 1823 (1) (January): 15–28. doi:10.1016/j.bbamcr.2011.06.007. http://www.ncbi.nlm.nih.gov/pubmed/21736903.

Bergh, O, K Y Børsheim, G Bratbak, and M Heldal. 1989. “High Abundance of Viruses Found in Aquatic Environments.” *Nature* 340 (6233) (August 10): 467–8. doi:10.1038/340467a0. http://www.ncbi.nlm.nih.gov/pubmed/2755508.

Bergthorsson, Ulfar, Dan I Andersson, and John R Roth. 2007. “Ohno’s Dilemma: Evolution of New Genes under Continuous Selection.” *Proceedings of the National Academy of Sciences of the United States of America* 104: 17004–17009. doi:10.1073/pnas.0707158104.

Blumberg, M.S., K. Deaver, and R.F. Kirby. 1999. “Leptin Disinhibits Nonshivering Thermogenesis in Infants after Maternal Separation.” *American Journal of Physiology* 276 (2): R606–R610. doi:10.1017/S0016756897007061.

Briglia, M, R I Eggen, D J Van Elsas, and W M De Vos. 1994. “Phylogenetic Evidence for Transfer of Pentachlorophenol-Mineralizing Rhodococcus Chlorophenolicus PCP-I(T) to the Genus Mycobacterium.” *International Journal of Systematic Bacteriology* 44 (3): 494–498. doi:10.1099/00207713-44-3-494.

Brüssow, Harald, Carlos Canchaya, and Wolf-Dietrich Hardt. 2004. “Phages and the Evolution of Bacterial Pathogens: From Genomic Rearrangements to Lysogenic Conversion.” *Microbiology and Molecular Biology Reviews : MMBR* 68 (3) (September): 560–602, table of contents. doi:10.1128/MMBR.68.3.560-602.2004. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=515249&tool=pmcentrez&rendertype=abstract.

Cai, Mian, and Luying Xun. 2002. “Organization and Regulation of Pentachlorophenol-Degrading Genes in Sphingobium Chlorophenolicum ATCC 39723.” *Journal of Bacteriology* 184 (17) (September): 4672–80. doi:10.1128/JB.184.17.4672. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=135293&tool=pmcentrez&rendertype=abstract.

Chaitra, M G, Sridhar Hariharaputran, Nagasuma R Chandra, M S Shaila, and R Nayak. 2005. “Defining Putative T Cell Epitopes from PE and PPE Families of Proteins of Mycobacterium Tuberculosis with Vaccine Potential.” *Vaccine* 23 (10) (January 26): 1265–72. doi:10.1016/j.vaccine.2004.08.046. http://www.ncbi.nlm.nih.gov/pubmed/15652669.

Cirelli, D P. 1978. “Patterns of Pentachlorophenol Usage in the United States of America - an Overview.” *Pentachlorophenol: Chemistry, Pharmacology, and Environmental Toxicology*: 13–18.

Clasen, Jessica L., Sean M. Brigden, Jerome P. Payet, and Curtis A. Suttle. 2008. “Evidence That Viral Abundance across Oceans and Lakes Is Driven by Different Biological Factors.” *Freshwater Biology* 53 (6) (June): 1090–1100. doi:10.1111/j.1365-2427.2008.01992.x. http://doi.wiley.com/10.1111/j.1365-2427.2008.01992.x.

Copley, Shelley D, Joseph Rokicki, Pernilla Turner, Hajnalka Daligault, Matt Nolan, and Miriam Land. 2012. “The Whole Genome Sequence of Sphingobium Chlorophenolicum L-1: Insights into the Evolution of the Pentachlorophenol Degradation Pathway.” *Genome Biology and Evolution* 4 (2) (January): 184–98. doi:10.1093/gbe/evr137. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3318906&tool=pmcentrez&rendertype=abstract.

Crawford, Ronald L, Carina M Jung, and Janice L Strap. 2007. “The Recent Evolution of Pentachlorophenol (PCP)-4-Monooxygenase (PcpB) and Associated Pathways for Bacterial Degradation of PCP.” *Biodegradation* 18 (5) (October): 525–39. doi:10.1007/s10532-006-9090-6. http://www.ncbi.nlm.nih.gov/pubmed/17123025.

Crooks, Gavin E., Gary Hon, John Marc Chandonia, and Steven E. Brenner. 2004. “WebLogo: A Sequence Logo Generator.” *Genome Research* 14: 1188–1190. doi:10.1101/gr.849004.

Cutting, W. C., H. G. Mehrtens, and M. L. Tainter. 1933. “Actions and Uses of Dinitrophenol.” *Journal of the American Medical Association* 101 (3) (July 15): 193. doi:10.1001/jama.1933.02740280013006. http://jama.jamanetwork.com/article.aspx?doi=10.1001/jama.1933.02740280013006.

Dai, Minghua, Julie Bull Rogers, Joseph R Warner, and Shelley D Copley. 2003. “A Previously Unrecognized Step in Pentachlorophenol Degradation in Sphingobium Chlorophenolicum Is Catalyzed by Tetrachlorobenzoquinone Reductase (PcpD).” *Journal of Bacteriology* 185 (1) (January 1): 302–10. doi:10.1128/JB.185.1.302-310.2003. http://jb.asm.org/cgi/doi/10.1128/JB.185.1.302-310.2003.

Darling, Aaron C E, Bob Mau, Frederick R Blattner, and Nicole T Perna. 2004. “Mauve: Multiple Alignment of Conserved Genomic Sequence with Rearrangements.” *Genome Research* 14 (7) (July): 1394–403. doi:10.1101/gr.2289704. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=442156&tool=pmcentrez&rendertype=abstract.

Darwin, Andrew J. 2005. “The Phage-Shock-Protein Response.” *Molecular Microbiology* 57 (3) (August): 621–8. doi:10.1111/j.1365-2958.2005.04694.x. http://www.ncbi.nlm.nih.gov/pubmed/16045608.

Das, Sarbashis, B.M. Fredrik Pettersson, Phani Rama Krishna Behra, Malavika Ramesh, Santanu Dasgupta, Alok Bhattacharya, and Leif A. Kirsebom. 2015. “Characterization of Three Mycobacterium Spp. with Potential Use in Bioremediation by Genome Sequencing and Comparative Genomics.” *Genome Biology and Evolution* 7 (7) (July): 1871–1886. doi:10.1093/gbe/evv111. http://gbe.oxfordjournals.org/lookup/doi/10.1093/gbe/evv111.

Edgehill, R U, and R K Finn. 1983. “Microbial Treatment of Soil to Remove Pentachlorophenol.” *Applied and Environmental Microbiology* 45 (3): 1122–5. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=242416&tool=pmcentrez&rendertype=abstract.

Endo, Ryo, Yoshiyuki Ohtsubo, Masataka Tsuda, and Yuji Nagata. 2007. “Identification and Characterization of Genes Encoding a Putative ABC-Type Transporter Essential for Utilization of Gamma-Hexachlorocyclohexane in Sphingobium Japonicum UT26.” *Journal of Bacteriology* 189 (10) (May): 3712–20. doi:10.1128/JB.01883-06.

http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1913331&tool=pmcentrez&rendertype=abstract.

Eppinger, Mark, Claudia Baar, Bodo Linz, Günter Raddatz, Christa Lanz, Heike Keller, Giovanna Morelli, Helga Gressmann, Mark Achtman, and Stephan C Schuster. 2006. "Who Ate Whom? Adaptive Helicobacter Genomic Changes That Accompanied a Host Jump from Early Humans to Large Felines." *PLoS Genetics* 2 (7) (July): e120. doi:10.1371/journal.pgen.0020120.eor. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1523251&tool=pmcentrez&rendertype=abstract.

Ezezika, Obidimma C., Sandra Haddad, Ellen L. Neidle, and Cory Momany. 2007. "Oligomerization of BenM, a LysR-Type Transcriptional Regulator: Structural Basis for the Aggregation of Proteins in This Family." *Acta Crystallographica Section F: Structural Biology and Crystallization Communications* 63 (5): 361–368. doi:10.1107/S1744309107019185.

Flynn, Julia M, Saskia B Neher, Yong In Kim, Robert T Sauer, and Tania A Baker. 2003. "Proteomic Discovery of Cellular Substrates of the ClpXP Protease Reveals Five Classes of ClpX-Recognition Signals." *Molecular Cell* 11 (3) (March): 671–83. http://www.ncbi.nlm.nih.gov/pubmed/12667450.

Gibson, Daniel G, Lei Young, Ray-yuan Chuang, J Craig Venter, Clyde A Hutchison Iii, Hamilton O Smith, and Nature America. 2009. "Enzymatic Assembly of DNA Molecules up to Several Hundred Kilobases." *Nature Methods* 6 (5): 12–16. doi:10.1038/NMETH.1318.

Habash, M B, L a Beaudette, M B Cassidy, K T Leung, T a Hoang, H J Vogel, J T Trevors, and H Lee. 2002. "Characterization of Tetrachlorohydroquinone Reductive Dehalogenase from Sphingomonas Sp. UG30." *Biochemical and Biophysical Research Communications* 299 (4) (December): 634–40. http://www.ncbi.nlm.nih.gov/pubmed/18042212.

Heidelberg, J F, J A Eisen, W C Nelson, R A Clayton, M L Gwinn, R J Dodson, D H Haft, et al. 2000. "DNA Sequence of Both Chromosomes of the Cholera Pathogen Vibrio Cholerae." *Nature* 406 (6795): 477–483. doi:10.1038/35020000.

Helmuth, R, and a Schroeter. 1994. "Molecular Typing Methods for S. Enteritidis." *International Journal of Food Microbiology* 21 (1-2) (January): 69–77. http://www.ncbi.nlm.nih.gov/pubmed/9809400.

Holden, Matthew T G, Richard W Titball, Sharon J Peacock, Ana M Cerdeño-Tárraga, Timothy Atkins, Lisa C Crossman, Tyrone Pitt, et al. 2004. "Genomic Plasticity of the Causative Agent of Melioidosis, Burkholderia Pseudomallei." *Proceedings of the National Academy of Sciences of the United States of America* 101 (39): 14240–14245. doi:10.1073/pnas.0403302101.

Hughes, A L. 1994. “The Evolution of Functionally Novel Proteins after Gene Duplication.” *Proceedings. Biological Sciences / The Royal Society* 256: 119–124. doi:10.1098/rspb.1994.0058.

Hyatt, Doug, Gwo-Liang Chen, Philip F Locascio, Miriam L Land, Frank W Larimer, and Loren J Hauser. 2010. “Prodigal: Prokaryotic Gene Recognition and Translation Initiation Site Identification.” *BMC Bioinformatics* 11 (January): 119. doi:10.1186/1471-2105-11-119. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2848648&tool=pmcentrez&rendertype=abstract.

Karn, Santosh Kr., S.K. Chakrabarty, and M.S. Reddy. 2010. “Pentachlorophenol Degradation by Pseudomonas Stutzeri CL7 in the Secondary Sludge of Pulp and Paper Mill.” *Journal of Environmental Sciences* 22 (10) (October): 1608–1612. doi:10.1016/S1001-0742(09)60296-5. http://linkinghub.elsevier.com/retrieve/pii/S1001074209602965.

Karn, Santosh Kr., S.K. Chakrabarty, and M. Sudhakara Reddy. 2010. “Characterization of Pentachlorophenol Degrading Bacillus Strains from Secondary Pulp-and-Paper-Industry Sludge.” *International Biodeterioration & Biodegradation* 64 (7) (October): 609–613. doi:10.1016/j.ibiod.2010.05.017. http://linkinghub.elsevier.com/retrieve/pii/S0964830510001356.

Kiefer, Philip M, Darla L McCarthy, and Shelley D Copley. 2002. “The Reaction Catalyzed by Tetrachlorohydroquinone Dehalogenase Does Not Involve Nucleophilic Aromatic Substitution.” *Biochemistry* 41 (4) (January): 1308–14. http://www.ncbi.nlm.nih.gov/pubmed/11802731.

Kim, Hong, Sun-Hyun Kim, Tae-Sun Shim, Mi-na Kim, Gill-Han Bai, Young-Gil Park, Sueng-Hyun Lee, et al. 2005. “Differentiation of Mycobacterium Species by Analysis of the Heat-Shock Protein 65 Gene (hsp65).” *International Journal of Systematic and Evolutionary Microbiology* 55 (Pt 4) (July): 1649–56. doi:10.1099/ijs.0.63553-0. http://www.ncbi.nlm.nih.gov/pubmed/16014496.

Kosuri, Sriram, Daniel B Goodman, Guillaume Cambray, Vivek K Mutalik, Yuan Gao, Adam P Arkin, Drew Endy, and George M Church. 2013. “Composability of Regulatory Sequences Controlling Transcription and Translation in Escherichia Coli.” *Proceedings of the National Academy of Sciences of the United States of America* 110: 14024–9. doi:10.1073/pnas.1301301110. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3752251&tool=pmcentrez&rendertype=abstract.

Krzywinski, Martin, Jacqueline Schein, Inanç Birol, Joseph Connors, Randy Gascoyne, Doug Horsman, Steven J Jones, and Marco A Marra. 2009. “Circos: An Information Aesthetic for Comparative Genomics.” *Genome Research* 19 (9) (September): 1639–45. doi:10.1101/gr.092759.109.

http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2752132&tool=pmcentrez&rendertype=abstract.

Letunic, Ivica, and Peer Bork. 2011. “Interactive Tree Of Life v2: Online Annotation and Display of Phylogenetic Trees Made Easy.” *Nucleic Acids Research* 39 (Web Server issue) (July): W475–8. doi:10.1093/nar/gkr201. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3125724&tool=pmcentrez&rendertype=abstract.

Liechti, George, and Joanna B Goldberg. 2012. “Helicobacter Pylori Relies Primarily on the Purine Salvage Pathway for Purine Nucleotide Biosynthesis.” *Journal of Bacteriology* 194 (4) (February): 839–54. doi:10.1128/JB.05757-11. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3272961&tool=pmcentrez&rendertype=abstract.

Machonkin, Timothy E, Patrick L Holland, Kristine N Smith, Justin S Liberman, Adriana Dinescu, Thomas R Cundari, and Sara S Rocks. 2010. “Determination of the Active Site of Sphingobium Chlorophenolicum 2,6-Dichlorohydroquinone Dioxygenase (PcpA).” *Journal of Biological Inorganic Chemistry : JBIC : A Publication of the Society of Biological Inorganic Chemistry* 15 (3) (March): 291–301. doi:10.1007/s00775-009-0602-9. http://www.ncbi.nlm.nih.gov/pubmed/19924449.

Maddocks, Sarah E, and Petra C F Oyston. 2008. “Structure and Function of the LysR-Type Transcriptional Regulator (LTTR) Family Proteins.” *Microbiology (Reading, England)* 154 (Pt 12) (December): 3609–23. doi:10.1099/mic.0.2008/022772-0. http://www.ncbi.nlm.nih.gov/pubmed/19047729.

Mane, S P, M G Dominguez-Bello, M J Blaser, B W Sobral, R Hontecillas, J Skoneczka, S K Mohapatra, et al. 2010. “Host-Interactive Genes in Amerindian Helicobacter Pylori Diverge from Their Old World Homologs and Mediate Inflammatory Responses.” *Journal of Bacteriology* 192 (12) (June): 3078–92. doi:10.1128/JB.00063-10. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2901691&tool=pmcentrez&rendertype=abstract.

Masters, Millicent, Garry Blakely, Andrew Coulson, Neil McLennan, Vollodymyr Yerko, and John Acord. 2009. “Protein Folding in Escherichia Coli: The Chaperonin GroE and Its Substrates.” *Research in Microbiology* 160 (4) (May): 267–77. doi:10.1016/j.resmic.2009.04.002. http://www.ncbi.nlm.nih.gov/pubmed/19393741.

McAllister, Kelly a., Hung Lee, and Jack T. Trevors. 1996. “Microbial Degradation of Pentachlorophenol.” *Biodegradation* 7 (1): 1–40. doi:10.1007/BF00056556.

Mohn, W W, and K J Kennedy. 1992. “Reductive Dehalogenation Of Chlorophenols By Desulfomonile-Tiedjei Dcb-1.” *Applied and Environmental Microbiology* 58 (4): 1367–1370.

Morano, Kevin A. 2007. “New Tricks for an Old Dog: The Evolving World of Hsp70.” *Annals of the New York Academy of Sciences* 1113 (October): 1–14. doi:10.1196/annals.1391.018. http://www.ncbi.nlm.nih.gov/pubmed/17513460.

Nagata, Yuji, Yoshiyuki Ohtsubo, Ryo Endo, Natsuko Ichikawa, Akiho Ankai, Akio Oguchi, Shigehiro Fukui, Nobuyuki Fujita, and Masataka Tsuda. 2010. “Complete Genome Sequence of the Representative {gamma}-Hexachlorocyclohexane-Degrading Bacterium Sphingobium Japonicum UT26.” *Journal of Bacteriology* 192 (21) (November): 5852–3. doi:10.1128/JB.00961-10. http://www.ncbi.nlm.nih.gov/pubmed/20817768.

Neher, Saskia B, Judit Villén, Elizabeth C Oakes, Corey E Bakalarski, Robert T Sauer, Steven P Gygi, and Tania A Baker. 2006. “Proteomic Profiling of ClpXP Substrates after DNA Damage Reveals Extensive Instability within SOS Regulon.” *Molecular Cell* 22 (2) (April 21): 193–204. doi:10.1016/j.molcel.2006.03.007. http://www.ncbi.nlm.nih.gov/pubmed/16630889.

Ning, Daliang, and Hui Wang. 2012. “Involvement of Cytochrome P450 in Pentachlorophenol Transformation in a White Rot Fungus Phanerochaete Chrysosporium.” Edited by Melanie R. Mormile. *PLoS ONE* 7 (9) (September 20): e45887. doi:10.1371/journal.pone.0045887. http://dx.plos.org/10.1371/journal.pone.0045887.

Nohynek, Liisa J., Eeva L. Suhonen, Eeva-Liisa Nurmiaho-Lassila, Jarkko Hantula, and Mirja Salkinoja-Salonen. 1995. “Description of Four Pentachlorophenol-Degrading Bacterial Strains as Sphingomonas Chlorophenolica Sp. Nov.” *Systematic and Applied Microbiology* 18 (4) (January): 527–538. doi:10.1016/S0723-2020(11)80413-3. http://linkinghub.elsevier.com/retrieve/pii/S0723202011804133.

Nordlie, K, and J Arthur. 1981. “Effect of Elevated Water Temperature on Insect Emergence in Outdoor Experimental Channels.” *Environmental Pollution Series A, Ecological and Biological* 25 (1) (May): 53–65. doi:10.1016/0143-1471(81)90114-8. http://linkinghub.elsevier.com/retrieve/pii/0143147181901148.

Ogle, James M, and V Ramakrishnan. 2005. “Structural Insights into Translational Fidelity.” *Annual Review of Biochemistry* 74 (January): 129–77. doi:10.1146/annurev.biochem.74.061903.155440. http://www.ncbi.nlm.nih.gov/pubmed/15952884.

Orser, C S, C C Lange, L Xun, T C Zahrt, and B J Schneider. 1993. “Cloning, Sequence Analysis, and Expression of the Flavobacterium Pentachlorophenol-4-Monooxygenase Gene in Escherichia Coli.” *Journal of Bacteriology* 175 (2) (January): 411–6. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=196155&tool=pmcentrez&rendertype=abstract.

Pignatello, J J, M M Martinson, J G Steiert, R E Carlson, and R L Crawford. 1983. “Biodegradation and Photolysis of Pentachlorophenol in Artificial Freshwater Streams.” *Applied and*

Environmental Microbiology 46 (5) (November): 1024–31. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=239514&tool=pmcentrez&rendertype=abstract.

Porrúa, Odil, Aroa López-Sánchez, Ana I Platero, Eduardo Santero, Victoria Shingler, and Fernando Govantes. 2013. “An A-Tract at the AtzR Binding Site Assists DNA Binding, Inducer-Dependent Repositioning and Transcriptional Activation of the PatzDEF Promoter.” *Molecular Microbiology* 90 (1) (October): 72–87. doi:10.1111/mmi.12346. http://www.ncbi.nlm.nih.gov/pubmed/23906008.

Ragan, Mark A, Timothy J Harlow, and Robert G Beiko. 2006. “Do Different Surrogate Methods Detect Lateral Genetic Transfer Events of Different Relative Ages?” *Trends in Microbiology* 14 (1) (January): 4–8. doi:10.1016/j.tim.2005.11.004. http://www.ncbi.nlm.nih.gov/pubmed/16303306.

Ricquier, D. 1999. “Mitochondrial Uncoupling Proteins.” *Current Opinion in Drug Discovery & Development* 2 (5): 497–504. http://www.ncbi.nlm.nih.gov/pubmed/19649977.

Robinson, James T, Helga Thorvaldsdóttir, Wendy Winckler, Mitchell Guttman, Eric S Lander, Gad Getz, and Jill P Mesirov. 2011. “Integrative Genomics Viewer.” *Nature Biotechnology* 29 (1) (January): 24–6. doi:10.1038/nbt.1754. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3346182&tool=pmcentrez&rendertype=abstract.

Rogall, T, J Wolters, T Flohr, and E C Böttger. 1990. “Towards a Phylogeny and Definition of Species at the Molecular Level within the Genus Mycobacterium.” *International Journal of Systematic Bacteriology* 40 (4) (October): 323–30. http://www.ncbi.nlm.nih.gov/pubmed/2275850.

Rokicki, Joe, David Knox, Robin D Dowell, and Shelley D Copley. 2014. “CodaChrome: A Tool for the Visualization of Proteome Conservation across All Fully Sequenced Bacterial Genomes.” *BMC Genomics* 15 (1): 65. doi:10.1186/1471-2164-15-65. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3908345&tool=pmcentrez&rendertype=abstract.

Rowe, Janet M, Jennifer M DeBruyn, Leo Poorvin, Gary R LeCleir, Zackary I Johnson, Erik R Zinser, and Steven W Wilhelm. 2012. “Viral and Bacterial Abundance and Production in the Western Pacific Ocean and the Relation to Other Oceanic Realms.” *FEMS Microbiology Ecology* 79 (2) (February): 359–70. doi:10.1111/j.1574-6941.2011.01223.x. http://www.ncbi.nlm.nih.gov/pubmed/22092569.

Ruangprasert, Ajchareeya, Sarah H. Craven, Ellen L. Neidle, and Cory Momany. 2010. “Full-Length Structures of BenM and Two Variants Reveal Different Oligomerization Schemes for

LysR-Type Transcriptional Regulators." *Journal of Molecular Biology* 404 (4): 568–586. doi:10.1016/j.jmb.2010.09.053. http://dx.doi.org/10.1016/j.jmb.2010.09.053.

Saber, D L, and R L Crawford. 1985. "Isolation and Characterization of Flavobacterium Strains That Degrade Pentachlorophenol." *Applied and Environmental Microbiology* 50 (6) (December): 1512–8. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=238790&tool=pmcentrez&rendertype=abstract.

Santos, SR, and Howard Ochman. 2004. "Identification and Phylogenetic Sorting of Bacterial Lineages with Universally Conserved Genes and Proteins." *Environmental Microbiology* 6: 754–759. doi:10.1111/j.1462-2920.2004.00617.x. http://onlinelibrary.wiley.com/doi/10.1111/j.1462-2920.2004.00617.x/full.

Sekirov, Inna, Shannon L Russell, L Caetano M Antunes, and B Brett Finlay. 2010. "Gut Microbiota in Health and Disease." *Physiological Reviews* 90 (3) (July): 859–904. doi:10.1152/physrev.00045.2009. http://physrev.physiology.org/content/90/3/859.short.

Sharma, Cynthia M, Steve Hoffmann, Fabien Darfeuille, Jérémy Reignier, Sven Findeiss, Alexandra Sittka, Sandrine Chabas, et al. 2010. "The Primary Transcriptome of the Major Human Pathogen Helicobacter Pylori." *Nature* 464 (7286) (March): 250–5. doi:10.1038/nature08756. http://www.ncbi.nlm.nih.gov/pubmed/20164839.

Shelton, D R, and J M Tiedje. 1984. "Isolation and Partial Characterization of Bacteria in an Anaerobic Consortium That Mineralizes 3-Chlorobenzoic Acid." *Applied and Environmental Microbiology* 48 (4): 840–8. doi:10.1016/j.plaphy.2004.10.010. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=241624&tool=pmcentrez&rendertype=abstract.

Slater, Steven C, Barry S Goldman, Brad Goodner, João C Setubal, Stephen K Farrand, Eugene W Nester, Thomas J Burr, et al. 2009. "Genome Sequences of Three Agrobacterium Biovars Help Elucidate the Evolution of Multichromosome Genomes in Bacteria." *Journal of Bacteriology* 191 (8) (April): 2501–11. doi:10.1128/JB.01779-08. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2668409&tool=pmcentrez&rendertype=abstract.

States, D, W Gish, and S Altschul. 1991. "Improved Sensitivity of Nucleic Acid Database Searches Using Application-Specific Scoring Matrices." *Methods* 3 (1) (August): 66–70. doi:10.1016/S1046-2023(05)80165-3. http://linkinghub.elsevier.com/retrieve/pii/S1046202305801653.

Tiirola, Marja A, Hong Wang, Lars Paulin, and Markku S Kulomaa. 2002. "Evidence for Natural Horizontal Transfer of the pcpB Gene in the Evolution of Polychlorophenol-Degrading Sphingomonads." *Society* 68 (9): 4495–4501. doi:10.1128/AEM.68.9.4495.

Tropel, David, and Jan Roelof van der Meer. 2004. “Bacterial Transcriptional Regulators for Degradation Pathways of Aromatic Compounds.” *Microbiology and Molecular Biology Reviews : MMBR* 68 (3) (September): 474–500, table of contents. doi:10.1128/MMBR.68.3.474-500.2004. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=515250&tool=pmcentrez&rendertype=abstract.

Union, International, O F Pure, and Applied Chemistry. 1987. “Environmental Chemistry of Pentachlorophenol.” *Pure and Appl. Chem.* 53: 1051–1080.

Van der Meer, Jan Roelof. 1997. “Evolution of Novel Metabolic Pathways for the Degradation of Chloroaromatic Compounds.” *Antonie van Leeuwenhoek* 71 (1-2): 159–78. doi:10.1023/A:1000166400935. http://link.springer.com/10.1023/A:1000166400935.

Vaz-moreira, Ivone, Olga C Nunes, Célia M Manaia, Ivone Vaz-moreira, Olga C Nunes, and M Manaia. 2011. “Diversity and Antibiotic Resistance Patterns of Sphingomonadaceae Isolates from Drinking Water Diversity and Antibiotic Resistance Patterns of Sphingomonadaceae Isolates from Drinking Water ” 77 (16). doi:10.1128/AEM.00579-11.

Warner, Joseph R, Sherry L Lawson, and Shelley D Copley. 2005. “A Mechanistic Investigation of the Thiol-Disulfide Exchange Step in the Reductive Dehalogenation Catalyzed by Tetrachlorohydroquinone Dehalogenase.” *Biochemistry* 44 (30) (August): 10360–8. doi:10.1021/bi050666b. http://www.ncbi.nlm.nih.gov/pubmed/16042413.

Xu, Ling, Katheryn Resing, Sherry L. Lawson, Patricia C. Babbitt, and Shelley D. Copley. 1999. “Evidence That pcpA Encodes 2,6-Dichlorohydroquinone Dioxygenase, the Ring Cleavage Enzyme Required for Pentachlorophenol Degradation in Sphingomonas Chlorophenolica Strain ATCC 39723 †.” *Biochemistry* 38 (24) (June): 7659–7669. doi:10.1021/bi990103y. http://pubs.acs.org/doi/abs/10.1021/bi990103y.

Xun, L, and C S Orser. 1991. “Purification and Properties of Pentachlorophenol Hydroxylase, a Flavoprotein from Flavobacterium Sp. Strain ATCC 39723.” *Journal of Bacteriology* 173 (14): 4447–53. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=208108&tool=pmcentrez&rendertype=abstract.

Xun, L Y, and C S Orser. 1991. “Purification of a Flavobacterium Pentachlorophenol-Induced Periplasmic Protein (PcpA) and Nucleotide Sequence of the Corresponding Gene (pcpA).” *Journal of Bacteriology* 173 (9) (May): 2920–6. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=207874&tool=pmcentrez&rendertype=abstract.

Xun, L., E. Topp, and C. S. Orser. 1992. “Purification and Characterization of a Tetrachloro-P-Hydroquinone Reductive Dehalogenase from a Flavobacterium Sp.” *Journal of Bacteriology* 174 (24): 8003–8007.

Young, Chun-Chun, Ian L Ross, and Michael W Heuzenroeder. 2012. “A New Methodology for Differentiation and Typing of Closely Related Salmonella Enterica Serovar Heidelberg Isolates.” *Current Microbiology* 65 (5) (November 14): 481–7. doi:10.1007/s00284-012-0179-3. http://www.ncbi.nlm.nih.gov/pubmed/22797864.

Zaninovich, Angel A., Marcela Raíces, Inés Rebagliati, Conrado Ricci, and Karl Hagmüller. 2002. “Brown Fat Thermogenesis in Cold-Acclimated Rats Is Not Abolished by the Suppression of Thyroid Function.” *American Journal of Physiology - Endocrinology And Metabolism* 283 (3) (September 1): E496–E502. doi:10.1152/ajpendo.00540.2001. http://ajpendo.physiology.org/lookup/doi/10.1152/ajpendo.00540.2001.

Zhou, D, and J Galán. 2001. “Salmonella Entry into Host Cells: The Work in Concert of Type III Secreted Effector Proteins.” *Microbes and Infection / Institut Pasteur* 3 (14-15): 1293–8. http://www.ncbi.nlm.nih.gov/pubmed/11755417.